Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Artificial Intelligence Ethics Canvas

A Tool for Ethical and Socially Responsible AI

Ankita Kalra

Supervised by Dr David Lewis

Co-Supervised by Harshvardhan Pandit

Declaration

I, hereby, declare that this project 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.

ANKITA KALRA

April 30, 2020

Abstract

This research conducted proposes a tool named AI Ethics Canvas which focuses on ensuring the socially responsible Artificial Intelligence by incorporating ISO 26000 guidelines in standardisation of AI. The aim of AI Ethics canvas tool is to bridge the gap between "AI expert", policy-makers and non-ethicists such as engineers, computer scientists by designing a tool which will consider ethical implications of AI and is easily integrated in an organisation/company/workspace.

An extensive review of related literature revealed that many international and national organisations are realising the need for standardisation of Artificial Intelligence Guidelines and have proposed various Documents. This is followed by proposing risks and mitigations of core issues established using ISO 26000. Core Issues of ISO 26000 were proposed as block headings for AI Ethics Canvas. In what follows next, is a critical analysis of various Documents proposed against ISO 26000.

In this context, the tasks of structuring ethical discussions about AI along with the business profit they generate is difficult due to a lack of methodologies. We present a way to relate the ethical and socially responsible considerations around AI with their business processes using the semantic web. This will allow an investigation of how ethics in AI, social responsibility in AI and business models affect each other and support a richer discussion around social responsibility. To assess the usefulness of AI Ethics Canvas, a questionnaire was circulated. The results suggest that it was moderately easy for students to understand the categories of AI Ethics Canvas and AI Ethics Canvas can be used as a tool for considering ethical and social responsible aspects of AI.

Acknowledgements

This MAI research project has been a tough and significant undertaking that I could not have accomplished without the amazing support and guidance of so many different people.

First and foremost, I am extremely thankful to **Dr David Lewis** and **Harshvardhan Pandit**, both of who have been an incredible support to me. Their insights, patience, guidance and support have been second-to-none. Whenever I was in need, they were there behind me.

I am extremely grateful and thankful to my loving family - Mom, Dad and Ayush. You have all invested so much in support of my dreams. I could never have accomplished any of this without you.

ANKITA KALRA

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This research seeks to explore the challenges which are being faced to ensure trust in Artificial Intelligence technologies by addressing social responsibility issues. With the rapid advancement in fields of Artificial Intelligence, there is always a "question mark" against the results which are generated by the algorithms. Since, it is not known how it is being generated what we are seeing, It becomes difficult to foresee the downsides of an AI technology which has resulted in many unprecedented situation and problems. This research has been undertaken to address these problems.

Section 1.1 gives an overview of the situations where Artificial Intelligent technologies were applied and have resulted in unforeseen and questionable results. It serves as a motivation for my research.

Section 1.2 defines the research question which is guided by the discussions provided in section 1.1.

Section 1.3 highlights the contribution made by this research to the domain of Artificial Intelligence.

Section 1.4 details the structure and content of the remaining report.

1.1 Motivation

The term Artificial Intelligence (AI) was first mentioned and introduced by John McCarthy in 1954. However AI has become more widely known to public audience during the past five to ten years. This is a very interesting fact and one of the reasons for taking this topic as the subject of my thesis. Yet as of now, Artificial intelligence is coined more and more by IT giants, private businesses and policymakers. The tasks which previously required human brains and intelligence are now being automated.The ability of AI to produce previously unattainable insights by integrating large streams of data has left tech giants amazed. But it is not all as convenient and good as it seems.

In parallel, many public concerns have been propelled to the forefront about the "trustworthiness" of the AI products.It is believed that such advanced applications raise concerning ethical issues because AI systems can reinforce what they have learned from the real-world data fed to their algorithms and might even amplify risks such as gender or racial bias. In fact, these systems are bound to make errors in decision when tested in an unfamiliar environment. And, since many of these AI programmed systems are "black boxes", it poses a significant challenge to provide a clear human-understandable explanation of how the results are generated.There have been a number of incidents reported in the past which emphasize this argument such as:

1. In January 2019, Google was fined approximately 50 million euros for "not disclosing how the data is collected across its services which includes Google Maps, YouTube, Search Engine to present personalized advertisements to its users".[34]

2. On March 2018, at around 10 P.M., Elaine Herzberg was wheeling her bicycle across a street in Arizona, when she was killed by a self-driving car.[38]

3. Facebook shares prices fell sharply in March 2018 because Cambridge Analytica, a political consultancy obtained personal information of 87 million users from Facebook.The then-presidential candidate Donald Trump who hired Cambridge Analytica, used this user data from Facebook to influence their behavior. [13]

4. The security and privacy concerns with voice assistants such as Amazon's Alexa are rising. Customers have reported that Alexa has been allegedly recording the previous

interactions. This has raised questions like "Should Alexa be allowed in your home?"
[12]

5. In 2014, Amazon developed a recruiting tool to select its team of software engineers,
   it might want to hire. It was later observed that the system began discriminating
   against women, so Amazon decided to abandon this in 2017.[36]

6. In 2016 Pro Publica built a commercially developed system that predicts the
   likelihood of the criminals to re-offend, thus helping judges to make better
   sentencing decisions, and later found out that the system was racist and
   discriminated against blacks.[26]

The controversies like these in the past have alerted policymakers and highlighted the
importance of ethically regulating the use of data with the consent of the user. This
predicament now also applies to the Artificial intelligence technologies which foresee an
even greater amount of data processing. So, it has highlighted the urgency to ensure new
practices are developed which will make sure that AI is operated alongside ethics in a
trustworthy manner without imposing risks to mankind. Mariya Gabriel, digital
commissioner quoted that "I am personally convinced that ethical guidelines will be
enablers of innovation for artificial intelligence". [25]

By ensuring that the AI technology which is being developed follows a set of certain
guidelines, some of the unforeseen results mentioned above can be avoided. Many
national and international organisations have established "a set of guidelines" which an AI
technology or product or system should abide by. The works done by the European Union
by laying down a set of guidelines named AI-HLEG are commendable in this regard. Also,
IEEE's Ethically Aligned Design which could be applied to Artificial Intelligence
technologies is noteworthy(It was designed for Autonomous/Intelligent Systems[A/IS]. AI
could be used synonomously here). Some other guidelines like Asilomar's and OECD have
structured arguments concerning AI, and made sure that there is no discrimination and
bias by the AI technologies. These authoritative works will be discussed in detail in Section
2.4.

Thus, exploring a set of guidelines that ensures that an AI technology does not discriminate
amongst mankind on any basis, is a key motivation behind this research. ISO 26000 is the

first international organisation that aims to establish a set of guidelines for ensuring trustworthiness in an AI technology. This work is still in its draft form. I have referred to the existing ISO 26000 guidelines which provides detailed explanation on integrating social responsibility in an organisation and proposed a set of guidelines which will aim to integrate social responsibility in an AI organisation.

1.2   Research Objective

The research question addressed by this research is as follows:

"How to ensure that Artificial Intelligence technology will execute socially responsible behavior towards the mankind? Is it possible to devise a tool which establishes and ensures socially responsible behavior and can be easily integrated in an ICT (Information and Communication Technology) sector?"

This question consists of 2 parts which form the key objective of my research that highlights:

1. The rapidly growing AI technology raises pressing questions about the accountability, ethics and impact of these technologies around the globe. This AI technology needs to be operated so that it does not bring havoc to the mankind. It can be achieved by ensuring that an AI technology operates in a socially responsible manner by considering the ethical impacts it foresees.

2. The second part of my research question focuses on the need for tools, which can be used in the ICT sector. The fact that there is a gap between the policy devising sector and the ICT sector, can not be ignored. We need to narrow the knowledge gap between AI experts and the people who use, interact with or are impacted by these AI technologies. Hence, there is a need for a tool, which can be easily integrated in ICT sector, ensures that the ICT sector is operated in a socially responsible manner and bridges the gap between them and the policy devising sector. Such tool will not only harness the full potential of AI systems, but will also ensure that they do not widen the existing inequalities and biases.

4

Keeping in mind the above research objective, my research is concentrated on the following points :

1. The efforts of various bodies to establish AI guidelines so that it does not result in any unprecedented consequences to mankind.

2. Aligning the AI guidelines to the established ISO 26000 guidelines which ensure that the AI organisation operates in a socially responsible behavior.

3. Proposing AI ethics canvas which serves as a tool for the ICT sector to consider integrating ethical impacts and socially responsible behavior when creating, deploying or using an AI technology.

4. Integrating AI Ethics Canvas tool with the Business Canvas and the Ethics Canvas using semantic web. This will ensure that this tool, when used in a workplace or organisation considers the business aspect, ethical implications and socially responsible behavior of an AI application.

5. Evaluating the user understanding and perceived usefulness of the AI Ethics Canvas tool when used for listing risks and mitigations of an AI application by using it in a classroom setting.

1.3   Contributions of Research

The final contributions of this research is a fully operational tool named AI Ethics Canvas which can be used to ensure that AI is operated in a socially responsible manner. The specific outcomes are as follow:

1. It serves as a research of how ISO 26000 guidelines can be aligned with organisations who are creating, deploying or using an AI technology by addressing various risks and mitigation.

2. It highlights the importance of AI ethics canvas tool in the ICT sector and exploits to serve it as a semantic web technology tool.

3. It serves as a survey report which comprises of opinions of people when asked to use ISO 26000 guidelines in terms of establishing socially responsible behavior of an AI technology.

1.4 Report Structure and Contents

The remainder of this dissertation is as follows:

Chapter 2 named as State of the Art aims to define the rhetoric efforts of various national and international organisations to consider "ethical impacts" of an AI technology by establishing a set of guidelines. These various efforts serve as a key motivation for this dissertation.

Chapter 3 named as Methodology discusses the ISO 26000 guidelines in detail, why I have used them as an inspiration for devising a set of guidelines which aims for establishing socially responsible behavior. It provides a detail of risks and mitigation of each of the core subject issue and a detailed comparison of ISO 26000 w.r.t other guidelines.

Chapter 4 named as Technical implementation which discusses the tool AI ethics canvas in detail. It aims to define the stages of completion of AI Ethics Canvas. It also establishes an AI ethics canvas tool as a semantic web technology by using it along with Business Canvas Model and Ethics Canvas Model. Towards the end, it establishes how the various controversies could have been avoided, or atleast predicted with the use of AI Ethics Canvas Tool.

Chapter 5 named as Evaluation provides a quantitative evaluation of the implemented guidelines. It also provides detail of the survey that were carried out amongst the students. The results are then evaluated and discussed.

Chapter 6 named as Conclusion. It provides an overall .,conclusion regarding the research as a whole and explores potential avenues for the continuation and enhancement of the research completed in this dissertation. It ends with final remarks based on the evaluation conducted.

Chapter 2

State of the Art

2.1 Introduction

This Chapter deals the state of the art in Artificial Intelligence and its guidelines with the aim of providing the reader with a comprehensive understanding of the research domain.

Section 2.2 provides a definition of Artificial Intelligence. It forms the basis for the user to understand the terminology

Section 2.3 highlights the need for building "standards" or a "set of guidelines" for AI to operate in a trustworthy manner

Section 2.4 lists the contributions by different technical giants in establishing guidelines for trustworthy AI in the first part. It also provides some controversies which involved technical giants when trying to establish Ethical AI. Further, it highlights the contributions towards Ethical AI by different organisations such as EU-HLEG, IEEE EAD, OECD, Asilomar. It also provides a timeline highlighting the various contributions towards the establishment of Ethical guidelines by various governmental, inter-governmental and multi-stakeholder organisations. This section ends with the discussion of how various of these guidelines are aligned with certain factors more than the others.

2.2   Definition of Artificial Intelligence

In the 21st Century, Artificial intelligence is coined more and more by IT giants, private businesses and policymakers. But what is it? There have been growing debates over the years about what exactly is Artificial Intelligence. As the technology advances, the definition of Artificial Intelligence changes. Every organisation be it multi-stakeholder, inter-governmental organisation or private sector, has emerged with a new definition for Artificial Intelligence. The European's commission HLEG on Artificial Intelligence offers a good place to start with:

"Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions." [10]

Aspects of this definition are relatable to those found in other works. Some of these works tend to define AI systems that act, some argue that these actions are merely general tasks that would require human brains otherwise. Other works establish AI by its known sub-fields like Machine Learning, deep neural networks, robotics or knowledge-based systems.

2.3   The need for standards in Artificial Intelligence

The last 5-10 years have seen some drastic developments in the field of Artificial Intelligence. Every week newspapers are filled with the information where a new record is set by clever algorithms of Artificial Intelligence. One of such events include when Google Deepmind's AlphaGo artificial intelligence system won against one of the world's top Go players, Lee sodol. [39] AlphaGo used general purpose learning and search algorithms and proved that machines could quickly develop superhuman capability. It is forecasted by

IDC, a premier global provider of market intelligence that 75% of commercial enterprise apps will use Artificial Intelligence by 2021.[15]

Like all new technologies there are risks and opportunities as well. French theorist Paul Virilio (1999, p.89) sums this up nicely with the following citation:

"When you invent the ship, you also invent the shipwreck; when you invent the plane you also invent the plane crash. Every technology carries its own negativity, which is invented at the same time as technological progress."

However, the risks and havoc brought by AI will be much worse than shipwreck or plane crash.

It is thus, fundamental to strike the right balance with AI in order for fears to remain fears. Using Ethical AI standards is one way of achieving this balance and is valuable in the following ways:

- Standards in AI are required since they will serve as a common terminology for use by the stakeholders involved.

- These standards will also enable detailed understanding, clear communication and fair decision making.

- It reduces the risks for new technologies such as that of AI.

- They also set a basis that abides by moral authority. The one who abuses them or ignores them at least feels a need to justify themselves and also their critics have a platform to cry foul.

Hence, ensuring Ethical AI standards will ensure higher level of adoption, use and interoperabitlity of AI technologies.

2.4 Contributions to Ethical AI by technical companies and various organisations

Due to the challenges mentioned in Section 1.1 that are coming forward, a huge debate has commenced around AI and ethics. One element of challenge is that who should devise, participate, contribute and determine the ethical guidelines for AI. Microsoft CEO Satya Nadella has said that the decision of idealising the guidelines should be left to tech companies such as the one he heads.[17] There is growing concern in the society that technical giants are having "leave it to us" attitude when it comes to writing the AI guidelines but dismissing them when convenient. Companies such as Google, Amazon, Apple, Facebook and Microsoft are amongst the leading companies using AI programmed machines, does that mean they get to decide and wrote the ethical rules for AI? But even within these technical giants, many problems are bound to occur. Google dissolved its AI ethics board merely 2 weeks after its foundation. [37] In fact, Google gave zero insight on how it elected its board members, which was criticised by many.

Many international bodies and national governments are trying to ensure the "trustworthiness" nature of Artificial Intelligence. They have undertaken social considerations of AI and tried to address issues like accountability, transparency, non-discrimination, privacy and security of data. Some have also proposed to consider established models of human rights as an underlying set of guidelines to design AI guidelines. These guidelines are not like the three Laws of Robotics by Issac Asimov [43] whose main focus was to prevent mankind from murderous robots. These guidelines aim to address problems that the society will face when integrating Artificial Intelligence into sectors like education, technology, mechanics, health care etc. For Example, If in future a person is rejected a loan application, these guidelines will make sure that the software did not discriminate based on your gender, that the software is able to explain the cause of rejection of application and it provides a detailed explanation what could the applicant do to get his application approved.

Figure 2.1 referenced from the document "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based approaches to Principles for AI" prepared by
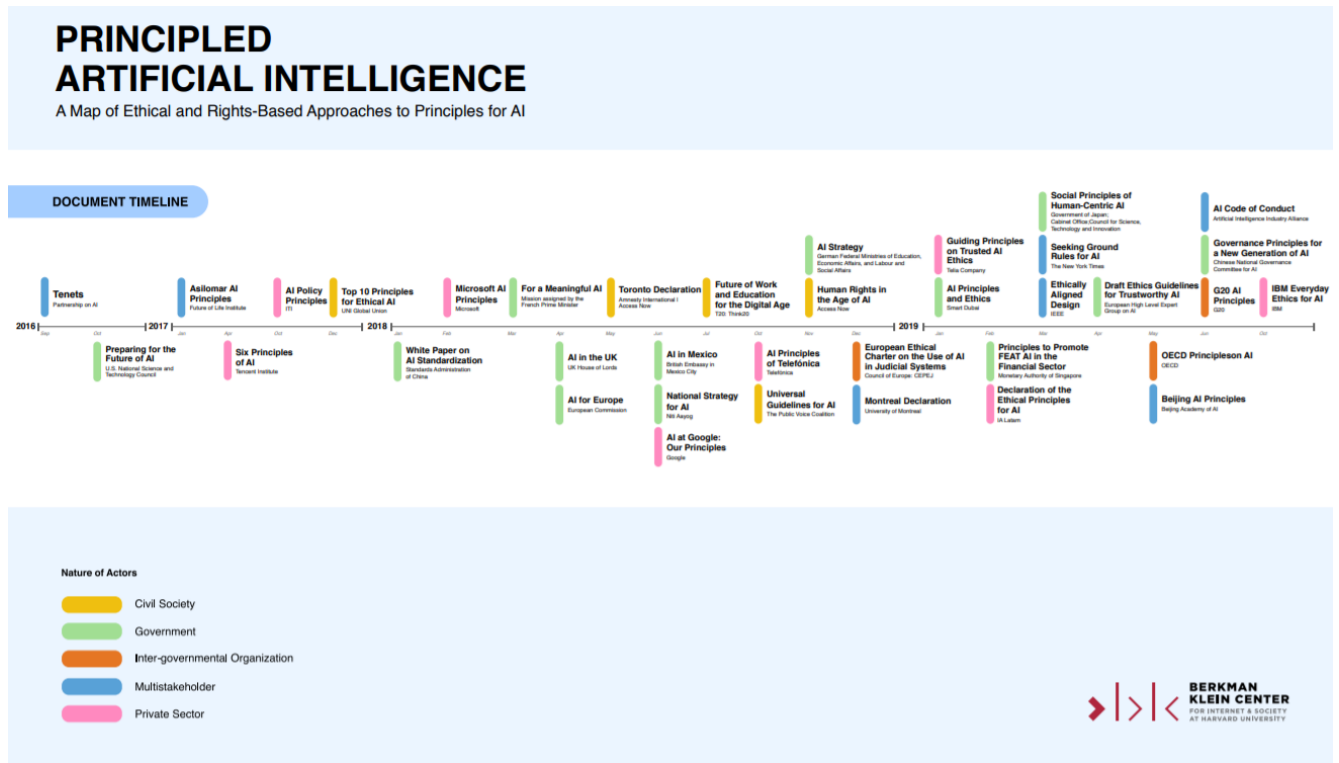
Figure 2.1: Timeline depicting various releases of documents concerning Artificial Intelligence

Berkman Klein Center provides a view of the timeline of releases of various documents. [11] Berkman Klein Center is a research center based at Harvard University which focuses on the study of Cyberspace. The timeline depicts that more and more organisations are undertaking the responsibility to establish a set of rules for the AI guidelines.

2019 has seen both EU's guidelines for Trustworthy AI termed as [HLEG-AI 2019] [9] and OECD (The Organisation for Economic Co-operation and Development) recommendations on Artificial Intelligence. EU constitutes of a governmental organisation whereas, OECD [24] constitutes an inter-governmental organisation 2.1. EU followed an open selection process and appointed 52 experts to a High-Level Expert Group from various sectors like academics, civil society, industry for forming guidelines on Artificial Intelligence. On the other hand, OECD set up a 50+ member expert group which consisted of representatives of 20 governments as well as leaders from various other sectors to scope a set of principles for AI. The adoption of the Principles for "Responsible Stewardship of Trustworthy AI" by the G20 (including China and Russia) in 2019 which also includes OECD

recommendations are a case in point.[29]

Professional initiatives by multi-stakeholder groups such as 2017 Asilomar principles on
Beneficial AI and 2019 IEEE's EAD(Ethically Aligned Design) [16] are complemented by
international works like UN's AI for good services, which aims to create a bridge between
trustworthy AI and Sustainable Development Goals. The Asilomar [33] Conference on
Beneficial AI, a conference organized by Future of Life Institute in 2017 which highlighted
23 guidelines for the research and development of AI. THE IEEE is taking a big step forward
by leading almost half a million technical professionals in 160 countries to address a vision
for prioritizing Human well-being with Autonomous and Intelligent Systems(A/IS). Note
that the IEEE principles analyse (A/IS) systems but artificial intelligence is used
synonymously here.

United Nations adopted "The 2030 Agenda for Sustainable development" [3] which
provided a shared blueprint for the establishment of people and the planet, for the present
and the future. The 17 Sustainable Development Goals (SDGs) and 169 targets which
demonstrate the scale and ambition of this new Agenda. It was adopted in 2015(The reason
it is not mentioned in the timeline in Figure 2.1). PwC proposed that AI is likely to
potentially contribute US$15.7 trillion to the global economy by 2030, which is the same
deadline UN proposed for Sustainable Development Goals"(SDGs).[23] Referring to the
paper [3] which establishes the role of AI in the achievement of the SDGs, it is stated that AI
can enable the accomplishment of 134 targets across all the goals, but it may also inhibit 59
targets. This proposal is also noteworthy in this regard. However, these principle led
models can present problems when applied to individual specific use cases. For example in
situations where the harm to a single individual needs to be weighed against the benefits of
a majority group. The document [20] by David Lewis et.al. proves validation to the above
argument. It states that the EU guidelines focus on the EU fundamental rights as grounding
principles, but is more aligned to those right without any referral to the right of solidarity,
which addresses the right to collective action in the workplace. The IEEE's Ethically Aligned
Design states that its grounding principles may be skewed towards the Western worldview
of ethics which are highly dominant in membership and accessibility to English-speaking
communities. However, it acknowledges that AI ethics may be established in different
moral worldviews such as Western Ethics, Buddhism, Ubuntu, Shinto.

Chapter 3

Methodology

3.1 Introduction

"Accountability!" Everyone has heard of this phrase when it comes to topics like self-driving cars or automated chatbots. It is used in various statements every day such as "Companies should be held accountable for the things they do.". Nevertheless, the everyday advancements in AI machines have opened various new challenges for companies with regard to their social responsibility. The companies which earlier had mottos like "move fast and break things" are choosing to move cautiously and hitting breaks before they make their next move. Thus everyone is realizing the need to embed terms like Accountability, Security, Privacy in their product releases. But when trying to embed these, companies are usually looking for some concrete set of rules which ensures that their product is socially responsible and is not harming the mankind in any way possible.

This Chapter provides an insight to various guidelines issued by the various organisations for ensuring social responsibility in AI. The ISO 26000 guidelines were chosen as a benchmark for the standardization. Further, there is a detailed comparison and mapping of ISO 26000 guidelines w.r.t other guidelines. This is done as a proof that most of they key issues of AI like **"accountability", "transparency", "security and privacy". and "non-discrimination"** are being addressed in the ISO 26000. This also provides the credibility of using ISO 26000 for AI standardization.

Section 3.2 provides an introduction to the various working groups under ISO. It also focuses on the task and issue division being handled by each of the groups under ISO/IEC JTC 1/SC 42.

Section 3.3 studies the reasons for undertaking ISO 26000 as a benchmark for establishing social responsibility in AI.

Section 3.4 provides the standard AI Guidelines. These guidelines are referred from ISO 26000 and modified for deploying, creating and using AI systems in organisations. Various questions are also formed which are listed as block questions in AI Ethics Canvas.

Section 3.5 provides mapping of various key issues extracted from guidelines w.r.t various organisations issuing the guidelines.

Section 3.6 provides mapping of various key issues extracted from guidelines w.r.t ISO 26000 guidelines. It also highlights the few occurrences in terms of important key issues.

Section 3.7 provides mapping and comparison of various guidelines w.r.t ISO 26000 with the aim to depict that most of the issues are being addressed in ISO 26000.

3.2   An Overview of ISO/IEC JTC 1/SC 42 Working

ISO/IEC JTC 1/SC 42 is the first international standards committee that is working for the establishment of guidelines for the entire AI ecosystem.[1] ISO/IEC JTC 1/SC 42 is a joint committee between ISO and IEC, the international standards development organizations(SDOs). In the creation of JTC 1/SC 42, JTC 1 has scoped SC 42 as a systems integration entity which will work with other ISO, IEC and JTC1 committees which are currently looking at the AI applications. JTC 1/SC 42 will [2]:

1. Serve as the focus and proponent for the JTC 1 AI standardization program.

2. Provide guidance to JTC 1, IEC, and ISO committees developing AI applications.

JTC 1/SC 42 currently has been divided in 5 Working Groups(WG):

- WG1 on Foundational Standards

- WG 2 on Big Data

- WG3 on AI Trustworthiness

- WG4 on gathering use cases and applications

- WG5 on AI computational methods

Amongst all of the these Working Groups, WG3 deals with the establishment of trustworthy AI. It evolved from a study group on AI Trustworthiness which was convened in Beijing inaugural plenary in 2018. WG3 works closely with all the other WGs.

SC 42 serves as a proponent for JTC 1 standardization program on AI by providing guidance to ISO, IEC and JTC 1 on regulating AI applications. SC 42 has been divided in the following groups which will cover the following aspects:

- Foundational Standards working group(WG1)

  With the increasing diversification of AI stakeholders, the need for foundational standards which can provide a common framework and vocabulary can not be ignored. WG1 will enable various AI stakeholders to talk the same language.

- Computational approaches and characteristics of artificial intelligence systems study group (SG 1)

  AI technologies are empowered by computational approaches and various algorithmic techniques. SG1 makes sure to standardize and allow best practices for innovation to occur.

- Trustworthiness study group (SG 2)

  SG 42 aims to ensure that AI technologies are trustworthy, robust, resilient, reliable, accurate and safe.

- Use cases and applications study group (SG 3)

  Use Cases are the currency by which standards development organizations collaborate with each other. SG 43 provides use cases for SC 42 to consider.

- Big data

    The JTC 1 big data program comes under SC 42.

Wael William Diab, Chair of SC 42 with more than 875 patents to his name in the field of information and communication technologies (ICT) highlights the approach of SC 42 in AI as[1]:

"One of the unique things about what IEC and ISO are doing through SC 42 is that we are looking at the entire ecosystem and not just one technical aspect. Combined with the breadth of application areas covered in IEC and ISO technical committees (TCs), this will provide a comprehensive approach to AI standardization with IT and domain experts."

An AI trustworthiness overview report (ISO/IEC DTR 24028) is under the ballot since 2019. Other reports on Trustworthiness of Neural networks (ISO/IEC WD TR 24029-1) as well as Bias in AI decision-making systems (ISO/IEC WD TR 24027) are on its way. WG 3 is performing an analysis of JTC1 and ISO and IEC standards regarding the gaps present in the documents.

In general, SC 42 is not aiming to invent brand new standards, instead they are examining the existing standards within ISO/IEC ecosystem and others, and mapping out specified areas w.r.t AI standardization. The main aim of the mapping is to identify gaps of existing AI standardisation to target specializations.

The new group of ethics and societal issues can consider ISO 26000 on social responsibility which lays down a standardised grounding for alignment with social responsibility issues. ISO Guide 82 lays down guidelines for integrating social responsibility issues in other standardization activities [ISO Guide 82]. ISO provides a set of 37 issues, classified under 6 core areas addressing social responsibility, which together with organisational governance can be used as a set of underlying principles for the formulation of new standards.

ISO guidelines can be used in any organizations regardless of size, type or location which is further elaborated in Section 3.3. These guidelines fundamentally aim to recognize social responsibility within the organization and also primarily focus on stakeholder identification and engagement. ISO guidance intends to establish social responsibility by

providing terms of 37 issues, each with expected behavior and suggested laws to address them. These issues are grouped under the following social responsibility core issues: Human Rights, Labour practices, the Environment, Fair operating practices, Consumer issues, and Community involvement and development. However ISO26000 does not specifically address the issues for the AI use, but they do provide an established set of issues that are sufficiently broad to provide a basis for mapping out specific ethical and societal issues associated with AI. Section 3.3 illustrates this argument by providing noteworthy points.

3.3   Considering ISO 26000 for standardizing AI Guidelines

In my thesis, I chose ISO 26000 as a standard when ensuring the trustworthy nature of AI machines in comparison to others. ISO 26000 Guidance on Social Responsibility which was launched by ISO (International Organization for Standardization) provides guidelines for establishing social responsibility(SR) behaviour named ISO 26000 or ISO SR. ISO consists of a network of national standards bodies representing from all regions of the world covering developed, developing and transitional economies comprising of more than 160 countries. Also the following factors are noteworthy for considering ISO 26000 to establish AI Guidelines:

- ISO 26000 is entirely a voluntary guidance standard. It contains no requirements such as the ones used when a standard is considered for "certification". There is no certification explicitly tied for using ISO. ISO recommends to its users to include the clause "used ISO 26000 as a guide to integrate social responsibility into our values and practices". [44]

- ISO 26000 has been designed for use by any type of organizations, be it hospitals, schools, small business groups charities (non-profit organisations). So far, many early users of ISO 26000 have been MNCs especially the ones based in Europe, Asia, particularly Japan. [18]

- ISO 26000 has been developed through a multi-stakeholder process,the meetings for which took place in 2005-2010 in 8 working Group sessions. 500 delegates (approx.)

participated in this process, who belonged to the six stakeholder groups: Industry, Government, NGO (non-governmental organization), Labour, Consumer, and SSRO (Service, Support, Research and Others - primarily academics and consultants). Leadership of task groups were divided between various "developed" and "developing" countries to enable various viewpoints from economic, cultural and social context. ISO group operates on the basis of consensus, so the finally agreed on standards were a result of various negotiations and deliberations. No single group was given the authority to block it, and not one single group was able to achieve it when others strongly disagreed. The fundamental aim of ISO 26000 was to make sure it is accessible and usable to all the organizations, in different countries, reflecting risks, aims and concerns of all the stakeholder groups.

- ISO 26000 reflects a wealth of international knowledge and experience since these are formed by a group of experts and esteemed delegations from all around the globe who need them and who know them. They aim to contribute to all the 3 dimensions of sustainable development- economic, social and environmental and form an international consensus from broad groups of stakeholders. Therefore, even though ISO 26000 are voluntary standards, they are widely adopted by business, government and society. This makes **ISO 26000 a strong candidate for laying down guidelines for AI addressing social responsibility issues.**

- ISO 26000 also includes recommendations from OECD(The organisation for Economic Cooperation and Development).Its fundamental principle is to promote policies that will improve the economic and social well-being of people around the world. It further provides a forum in which governments can work together to share experiences and seek solutions to common problems. [19]

All in all, ISO26000 is more explicit on the principles of "accountability", "transparency", "ethical behaviour" and "respect for stakeholder interests" which is a case in the point. Applying and promoting these principles lays down a concrete set of guidelines for Artificial Intelligence. These various factors make ISO 26000 as an inspiration when forming AI guidelines.

3.4 Core Issues of ISO 26000-Risks and Mitigations

In this section, I propose to consider ISO 26000 for forming AI standards. ISO 26000 provides a set of 37 issues, grouped under 6 core areas of social responsibility, which together with issues of organizational governance can be used as a form of checklist to consider in formulating new standards.

Further in this section I propose to consider mapping of ethical and societal AI issues against the social responsibility core issues with an identification of: **risks** for each core issue(which were mapped from the ISO 26000 **Description of each issue**) and **mitigation** for each core issue(which were mapped from the ISO 26000 **Related actions and expectations**). At the end of each core area, there will be questions addressed in that context. These questions will be incorporated in AI Ethics canvas (Figure4.2). These questions are formed on the basis of mitigations of each core issue under that core area, thereby, addressing these questions will help in preventing, predicting and foreseeing the risks for each issue. This will ensure socially responsible behaviour of an AI technology.(Note that the risks and mitigations mapped in this section are from the text in the ISO 26000 Guidelines. The link to the guide is not cited since it is not available on the internet. One has to request a copy for the same.)

Below is a mapping of ethical and societal AI issues against the social responsibility core issues alongside with an identification of risks, mitigations and questions addressed:

3.4.1 Organisational Governance

Questions addressing Organisational Governance: Does the organisation owning the product takes full responsibility of any damage reported to the society by its use? Does the organisation have made structured laws available which apply social responsibility?(Note that Organisational Governance is not a core area but has still been included in AI Ethics Canvas Design)

19

3.4.2 Human Rights

3.4.2.1 Issue 1: Due Diligence

Risks: How will the organisation which tends to deploy, use or create AI system deal with a human rights violation? How will the stakeholders of the AI system will believe that it is not going to violate their basic rights?

Mitigation: An organisation which tends to create, deploy or use AI systems can ensure to formulate a meaningful human rights policy through which users of the AI system will be well aware of the impacts on their human rights. It will help in establishing **accountability** for the AI system.

3.4.2.2 Issue 2: Human rights risk situation

Risks: What if an organisation which tends to deploy, use or create AI system influences criminal behavior online? What if the AI system indulges in activities that can affect or involve children or results in a culture of corruption? What if the AI system results in a biased judicial system? If an AI system such as a chatbot is developed which sends out fake news thereby resulting in political instability, how will it be mitigated?

Mitigation: These situations may require an enhanced process of due diligence to ensure respect for human rights. An organisation which tends to create, deploy or use AI systems can ensure to formulate an independent human rights impact assessment. In responding, an organization should consider the potential consequences of its actions so that the desired objective of respecting human rights is actually achieved. This will help in ensuring **peace, justice and development of strong institutions**.

3.4.2.3 Issue 3: Avoidance of Complicity

Risks: How will an organisation tend to deploy, use or create AI system deals with a situation where it finds itself benefiting economically from supplier's abuse of fundamental rights to work? What if an organization is tolerating action by security forces to suppress a

peaceful protest against its decisions and activities? What if the AI system tends to discriminate in employment laws against particular groups?

Mitigation: An organization which tends to create, deploy or use AI systems should verify that its security arrangements respect human rights and are consistent with international norms and standards for law enforcement. An organisation should not provide goods or services to an entity that uses them to carry out human rights abuses or enter into a formal or informal partnership with them. This will help in ensuring the key principle of **transparency, openness and accountability**.

3.4.2.4 Issue 4: Resolving Grievances

Risks: What if an AI machine algorithmically refuses a loan or an insurance and discriminates based on race, gender, sexuality etc.? Is there any mechanism ensured by the organisation which tends to create, deploy or use AI systems if human rights have been abused? Is it possible to inform the organisation and seek redress?

Mitigation: An organization which tends to create, deploy or use AI systems should establish, or otherwise ensure the availability of remedy mechanisms for its own use and that of its stakeholders. These mechanisms can be made effective by exhibiting: clear, transparent and sufficiently independent governance, accessible regardless of language, lack of awareness, distance, disability or fear of reprisal, predictability in processes and outcomes, equity amongst parties, openness to external scrutiny and opportunity to reach resolutions through dialogue and mediation. These mechanisms will help in establishing **accountability** of AI system creation and operation.

3.4.2.5 Issue 5: Discrimination and vulnerable groups

Risks: How can an organisation which tends to create, deploy or use AI systems avoid discrimination of vulnerable groups? How can it ensure equal opportunity and respect for all individuals?

Mitigation: An organization which tends to create, deploy or use AI systems should

examine its own operations and the operations of other parties within its sphere of influence to determine whether direct or indirect discrimination is present. It should also ensure that it is not contributing to discriminatory practices through the relationships connected to its activities. If this is the case an organization should encourage and assist other parties in their responsibility to prevent discrimination. If this is not successful it should reconsider its relations with such organizations. It may, for example, undertake an analysis of typical ways in which it interacts with women, as compared with men, and consider whether policies and decisions in this regard are objective or reflect stereotyped preconceptions. It may wish to seek advice from local or international organizations with expertise in human rights. An organization may be guided by the findings and recommendations of international or national monitoring or investigative procedures. This will help in ensuring the key principle of **No Poverty, Inclusion, decent work,non-discrimination and justice**.

3.4.2.6   Issue 6: Civil and Political Rights

Risks: If an organisation which tends to create, deploy or use AI systems finds itself amidst in a situation where the absolute rights like the right to a life with dignity, the right to freedom from torture, the right to security of person, the right to own property, liberty and integrity of the person, and the right to due process of law and a fair hearing when facing criminal charges are being violated? What if an organisation tends to exploit the location of a person via its AI system or is involved in altering the digital content displayed to the user by AI system?

Mitigations: An organization which tends to create, deploy or use AI systems should respect all individual civil and political rights. An organization should not aim to suppress anyone's views or opinions, even when the person expresses criticism of the organization internally or externally. An organisation can take the course of due diligence amongst its stakeholders to ensure that no such right is violated. This will aim to the fulfillment of principles like **No poverty, industry innovation, peace, fairness, justice and non-discrimination.**

3.4.2.7   Issue 7: Economic, social and cultural rights

Risks: If an organisation which tends to create, deploy or use AI systems finds itself amidst in a situation where the right to: education; work in just and favourable conditions; freedom of association; an adequate standard of health; a standard of living adequate for the physical and mental health and well-being of himself or herself and his or her family is being violated?

Mitigations: An organization can ensure measures and include due diligence to ensure AI systems don't infringe on these rights. E.g. through labor replacement, and use of AI system to fulfill such rights, E.g.by facilitating affordable education and lifelong learning and access to AI services by poor people. Organizations should, where appropriate, consider adopting or maintaining specific policies to ensure the efficient distribution of essential goods and services where this distribution is endangered.It will help in promoting principles like **No Poverty, Sustainable development, peace, fairness, non-discrimination**.

3.4.2.8   Issue 8: Fundamental principles and rights to work

Risks: What if an organisation which tends to create, deploy or use AI systems discriminates amongst the people when hiring, or employing, or nominating? Is the organisation taking any measures to ensure the elimination of all forms of forced or compulsory labour?

Mitigations: An organization who intends to create, deploy or use AI can undertake to ensure that employment policies free from discrimination and zero engagement in benefit from use of forced or compulsory labour. This can be ensured in algorithmic selection for recruiting and in algorithmic work performance assessment. Another measure can be ensuring the freedom of association and collective bargaining of workers in AI-mediated businesses. Organizations should also take steps to prevent harassment in the workplace by regularly assessing the impact of its policies and activities on promotion of equal opportunities and non-discrimination. Actions like establishing workplaces for persons with disabilities to help them earn a living under suitable conditions, and establishing or

participating in programs that address issues such as promotion of employment for youth and older workers, equal employment opportunities for women and more balanced representation of women in senior positions will also help in ensuring **Decent Work and Economic growth** amidst the organisation.

Questions addressing Human Rights: How the product will affect basic human rights? Are there enough guidelines available for the product how it may be used, where it may be used? Does it violate any political, social and cultural rights of the stakeholders involved?

3.4.3 Labor Practices

3.4.3.1 Issue 1: Employment and employment relationships

Risks: What measures are adopted by an organisation which tends to create, deploy or use AI systems for the people working to always receive the protections and rights that they are entitled to receive? Does the organisation imposes obligations on both employers and employees in the interest of both the organization and society?

Mitigations: An organisation which tends to create, deploy or use AI systems should ensure equal opportunities for all workers and not discriminate either directly or indirectly in any labour practice. It can be achieved by providing reasonable notice, timely information and, jointly with worker representatives where they exist, considering how to mitigate adverse impacts to the greatest possible extent when considering changes in its operations, such as closures that affect employment. An organization can also seek assurances that these issues are observed in partner organizations such as those providing AI components or training data or using provided AI-based services. This will ensure **decent work and economic growth, practices of fairness and non-discrimination** in the workplace

3.4.3.2 Issue 2: Conditions of work and social protection

Risks: Does an organisation which tends to create, deploy or use AI systems have ensures policies and practices to mitigate the reduction or loss of income in case of employment injury, illness, maternity, parenthood, old age, unemployment, disability or financial

hardship and to provide medical care and family benefit? Is the organisation successfully able to provide compensation, working time, rest periods, holidays, disciplinary and dismissal practices to the people working?

Mitigations: An organization that employs people that are impacted by its creation, deployment or use of AI systems can undertake to provide decent conditions of work complying with regards to wages, weekly rest, holidays health and safety and ability to combine work with family responsibilities. It can compensate workers for overtime in accordance with laws, regulations or collective agreements. It will help in achieving **decent work and economic growth, practices of fairness and non- discrimination** in the organisation.

3.4.3.3 Issue 3: Social dialogue

Risks: What measures are ensured by organisation which tends to create, deploy or use AI system for allowing the exchange of information between or among its employers or workers?

Mitigations: An organisation which tends to create, deploy or use AI system should respect at all times the right of workers to form or join their own organizations, groups to advance their interests or to bargain collectively. Such groups could include new categories which emerge in AI-based sectors like content annotators, creator. This can be achieved by adopting framework agreements supplemented by local organization-level agreements in accordance with national law or practice.

3.4.3.4 Issue 4: Health and safety at work

Risks: How will an organisation which tends to create, deploy or use AI system ensures health and safety of people such as robot co-workers, or semi-autonomous vehicle operators?

Mitigations: An organisation which tends to create, deploy or use AI system can record or investigate all health and safety incidents and problems in order to minimize or eliminate

them. It can provide adequate training on the functioning of AI systems and provide equal health and safety protection on all relevant matters. It can also develop, implement and maintain an occupational health and safety policy based on the principle that strong safety and health standards and organizational performance are mutually supportive and reinforcing. It will help in promoting **decent work and economic growth** within the organisation.

3.4.3.5   Issue 5: Human development and training in the workplace

Risks: Does organisation which tends to create, deploy or use AI system ensures any measures for the people who are unemployed due to the replacement of their work roles by AI systems?

Mitigations: An organization that employs people that are impacted by its creation, deployment or use of AI systems can undertake to further human development and training as work changes due to the introduction of AI systems, including development and training to address employment lost and deskilling of work roles arising from AI automation of worker tasks. It can also establish joint labour-management programmes that promote health and well-being. This will ensure **Quality Education, decent work and Economic Growth and promotion of Innovation**.

Questions addressing Labor Practices: Could the product put people out of work? Does the product promote human development and training in the work place? How the product could affect employment relationships? Does the product promote health and safety at work?

3.4.4   The Environment

3.4.4.1   Issue 1: Prevention of pollution

Risks: How will an organisation which tends to create, deploy or use AI system ensure pollution included emissions to air, discharge to water etc is prevented?

Mitigations: An organisation which tends to create, deploy or use AI can measure, record and report on its significant sources of pollution and reduction of pollution, water consumption, waste generation and energy consumption. It can also engage with local communities regarding actual and potential polluting emissions and waste, related health risks, and actual and proposed mitigation measures. This will help in promoting Sustainable Development Goals such as **Clean Water Sanitation, Clean Energy, Climate Action and Societal Well-being.**

3.4.4.2   Issue 2: Sustainable resource use

Risks: How will an organisation which tends to create, deploy or use AI system ensure sustainable use of resources?

Mitigations: An organization that intends to create, deploy or use AI systems can implement resource efficiency measures to reduce its use of energy, water and other resources, considering best practice indicators and other benchmarks. It can complement or replace non-renewable resources where possible and use recycled materials and reuse materials as much as possible. This will help in promoting Sustainable Development Goals such as **Quality Education, Sustainable Development, Decent Work and Economic Growth.**

3.4.4.3   Issue 3: Climate change mitigation and adaptation

Risks: How will an organisation which tends to create, deploy or use AI system ensures the impact of their activities on climate change e.g., through cloud-based AI?

Mitigations: An organisation which tends to create, deploy or use AI system should mitigate climate change impacts by identifying the sources of direct and indirect emissions, promote purchasing of energy-efficient goods and development of energy-efficient products and services. This will help in promoting Sustainable Development Goals such as **Zero Hunger, Sustainable Development.**

3.4.4.4 Issue 4: Protection of the environment, biodiversity and restoration of natural habitats

Risks: How will an organisation which tends to create, deploy or use AI system ensures impact of their activities on biodiversity and ecosystem services?

Mitigations: An organization that intends to create, deploy or use AI systems should give highest priority to avoiding the loss of natural ecosystems, second to restoring ecosystems, and finally, if the former two actions are not possible or fully effective, to compensating for losses through actions that will lead to a net gain in ecosystem services over time. Not only that AI system may contribute to the protection of ecosystems through its algorithmic application to sense, model and predict environmental issues. This will help in promoting Sustainable Development Goals such as **Zero Hunger, Sustainable Development, Climate Action, Life, Environmental and Social Well-being of individuals.**

Questions addressing The Environment: Is the product consuming resources by sacrificing the need of the future generation? Does the product displaces the activities by its user and stakeholders that incur higher or lower emissions? What happens with the product's use of resources, energy?

3.4.5 Fair Operating Practices

3.4.5.1 Issue 1: Anti-corruption

Risks: Is the organisation which tends to create, deploy or use AI system lays down any mechanism to curb corruption?

Mitigations: AI being used in sectors like finance, investments, accounts can foresee practices of corruption. An organisation should establish and maintain an effective system to counter corruption by ensuring to bring violations of the criminal law to the attention of appropriate law enforcement authorities. This will establish **Peace, Justice and Privacy, Data Governance.**

3.4.5.2   Issue 2: Responsible political involvement

Risks: What if the organisation which tends to create, deploy or use AI system finds itself amidst of situation where it is influencing political processes by generating fake news or manipulating users content?

Mitigations: An organisation that tends to create, deploy or use AI system should be transparent regarding its policies and activities related to political contributors and political involvement. It can establish and implement policies and guidelines to manage the activities of people and train its employees and representatives to raise their awareness regarding responsible political involvement. This will establish a strong foundation for ensuring **non-discrimination, fairness and accountability.**

3.4.5.3   Issue 3: Fair competition

Risks: How will an organisation which tends to create, deploy or use AI system deal with situations like price-fixing, bid rigging, competitive bidding?

Mitigations: An organisation should establish procedures and other safeguards to prevent engaging in or being complicit in anti-competitive behaviour and conduct its activities in a manner consistent with competition laws and regulations.

3.4.5.4   Issue 4: Promoting social responsibility in the value chain

Risks: What if the organisation which tends to create, deploy or use AI system encourage socially responsible behaviour in the value chain? Does it influence other organisations to promote, adopt and support social responsibility behavior?

Mitigations: An organisation can promote socially responsible behavior by following adequate purchasing practices, such as ensuring that fair prices are paid and that there are adequate delivery times and stable contracts. Also carrying out appropriate due diligence and monitoring of the organizations with which it has relationships, with a view to preventing compromise of the organization's commitments to social responsibility will be

helpful. This will help in ensuring **Decent Work, Economic Growth, Innovation, Responsible Consumption, Transparency and accountability**

3.4.5.5   Issue 5: Respect for property rights

Risks: Does the organisation which tends to create, deploy or use AI system discriminates algorithmically amongst individuals when granting them a physical property or an intellectual property? How does it mitigate such risks?

Mitigations: An organisation should conduct proper investigations to be confident it has lawful title permitting use or disposal of property. It should not engage in activities that violate property rights, including misuse of a dominant position. It will ensure the **privacy protection of data, transparency in organisation's methods and accountability of its actions**.

Questions addressing Fair Operating Practices: Could the product affect democratic decision making in an opaque manner? Are people contributing to data driving AI system in a fair and transparent manner? Does the product ensures fair competition amongst people using it or not using it? Does the product at present/later stage encourages political involvement? Does the usage of product amongst people results in corruption?

3.4.6   Consumer Issues

3.4.6.1   Issue 1:Fair marketing, factual and unbiased information and fair contractual practices

Risks: Is the organisation providing the complete information of AI product to its stakeholders for ensuring their trust? If so, how?

Mitigations: An organisation should ensure it does not provide misleading, deceptive, fraudulent information by consenting to share relevant information in a transparent manner. This could also be achieved by enforcing clear legible and understandable language in the contracts. It will ensure **Technical Robustness, Transparency,**

Accountability, Non-discrimination and Inclusion in the organisation.

3.4.6.2   Issue 2: Protecting consumers' health and safety

Risks: How does an organisation ensure that it is not harming consumers and ensures complete protection of them? What is a self autonomous car ends up killing people? What are the measures taken to avoid that?

Mitigations: An organisation can mitigate the above risks by performing a human health risk assessment of AI products and services before the introduction of new materials. Also once the product is placed on the market and presents an unforeseen hazard, organisation should stop the services or withdraw all AI products that are still in the distribution chain. It will promote the principles of **Responsible consumption, Technical Robustness, Privacy Protection, Environmental, Social Well being** and establish **accountability of the organisation.**

3.4.6.3   Issue 3: Sustainable consumption

Risks: Does the organisation acknowledge the adverse effects of producing AI products on the environment? Does it ensures the necessary steps to reduce them?

Mitigations: An organisation should design sustainable products and packaging which can be easily used, reused, repaired and recycled. It should also offer transparent information to consumers about energy and greenhouse gas emissions resulting across the value chain for AI-based service use to enable them to make better decisions on the sustainability of their consumption. It will ensure **sustainable development principles, Transparency and accountability** of the organisation.

3.4.6.4   Issue 4: Consumer service, support, and complaint and dispute resolution

Risks: An AI product produced by the organisation is definitely going to be a little expensive than the normal products. What if, a consumer is not satisfied with it? Or what if the user wished to return it?

Mitigations: An organisation should offer maintenance and repair at a reasonable price and at accessible locations and make information readily accessible on the expected availability of spare parts for AI products. If appropriate, it should offer warranties that exceed periods of AI products guaranteed by law. It will promote principles like **innovation, transparency, social well-being and accountability**.

3.4.6.5 Issue 5: Consumer data protection and privacy

Risks: How are the stakeholders of the AI product made aware that their data is secured and not being misused?

Mitigations:An organisation should minimize the collection of personal data. It should ensure to employ mechanism to implement and demonstrate compliance to relevant privacy and data protection regulations. It should be transparent on the data held on an individual and should offer mechanisms for resolving disputes concerning these features. It will establish **Privacy protection, transparency, accountability and explainability.**

3.4.6.6 Issue 6: Access to essential services

Risks: How does an organisation ensure access to essential services by using an AI product?

Mitigations: An organization that intends to create, deploy or use AI systems that involve collection and processing of data can offer mechanisms for transparent, explainable and non-discriminatory consumer access to essential services, including access to health and wellbeing information and data. AI products will also play a fundamental role in the fulfillment of essential services like accessing applications for health insurance. This will lay a foundation for the principles such as **Non-discrimination, Accountability and Inclusion**.

3.4.6.7 Issue 7: Education and awareness

Risks: How does an organisation create awareness and education about the use of AI products?

Mitigations: The organisation should address information about risks related to use and any necessary precaution. It should also address information on appropriate laws and regulations, ways of obtaining redress and agencies and organizations for consumer protection. This will establish **Sustainable Development, Technical Robustness and Accountability** in and around the organisation.

Questions addressing Consumer Issues: Can the AI System explain its decision to the affected consumers in a transparent way? Does the product in ant way violate consumer's privacy by exposing his data? Is there enough explainability about the product as to why certain data is being asked for? Are essential services made available to the consumers about how will there data be used and disposed off?

3.4.7 Community Involvement and Development

3.4.7.1 Issue 1:Community involvement

Risks: Does the organisation which intends to create, deploy or use AI products has proactive outreach to its community? Is it aware for the community needs and priorities and take responsibility of its impacts on society and the environment?

Mitigations: An organisation should participate in local associations ad possible and appropriate and encourage in maintaining a transparent relationship with local government officials and political representatives. It can proactively reach out to the community by contributing to policy formation and the establishment implementation, monitoring and evaluation of the development programs. It will ensure **transparency of organisations procedures and accountability for its actions**.

3.4.7.2 Issue 2: Education and culture

Risks: Does the organisation which intends to create, deploy or use AI products preserve and promote education and culture with respect for human rights?

Mitigations: The organisation which intends to create, deploy or use AI products should

encourage the enrolment of children in formal education about AI products and promote learning opportunities for vulnerable and discriminated groups. It should also help to conserve and protect cultural heritage and where appropriate, promote the use of traditional knowledge and AI technologies of indigenous communities. This will help in ensuring **Quality education amongst the community**.

3.4.7.3 Issue 3: Employment creation and skills development

Risks: Does an organisation which intends to create, deploy or use AI products ensure employment promotion and assist people to secure decent and product jobs?

Mitigations: An organisation should select technologies that maximize employment opportunities. It should analyse the impact of its investment decisions on employment creation and where economically viable, should consider making direct investments that alleviate poverty. It can also consider participating in local and national skills development programs including apprenticeship programmes which will help to resolve the issues of labor displacement created by the AI technologies. It will help to promote **sustainable development goals and establishing the accountability of the organisation using AI products.**

3.4.7.4 Issue 4: Technology development and access

Risks: Does the organisation which intends to create, deploy or use AI technologies improve access to its technologies? If so, how?

Mitigations: An organisation should consider developing local and traditional knowledge and should consider engaging in partnerships with organizations, such as universities or research laboratories, to enhance AI technology development. It will aim at supporting non-discrimination amongst community, accountability of organisation and promote partnership for the Goals.

3.4.7.5 Issue 5: Wealth and income creation

Risks:r Does the organisation ensure to create an environment that support entrepreneurship opportunities? How does it ensure to play a positive role in the development of the community?

Mitigations: An organisation should consider giving preference to local suppliers of products and services and contribute to local development where possible. It should undertake initiatives to strengthen the ability and opportunities for locally based suppliers. It should also consider contributing to durable programs and partnerships that assist community members especially women. It will promote principles of **non-discrimination, transparency to procedures and accountability of the organisation**.

3.4.7.6 Issue 6: Health

2. Does the organisation which intends to create, deploy or use AI products preserve and promote education and culture with respect for human rights?

Mitigations: The organisation which intends to create, deploy or use AI products should encourage the enrolment of children in formal education about AI products and promote learning opportunities for vulnerable and discriminated groups. It should also help to conserve and protect cultural heritage and where appropriate, promote the use of traditional knowledge and AI technologies of indigenous communities. This will help in ensuring **Quality education amongst the community**.

Risks: Does an organisation respects the right to health and contribute to the promotion of health, and prevention of health threats and diseases?

Mitigations: An organisation should consider raising awareness about major health threats such as HIV/AIDS, and support long-lasting access to essential health care services. It should consider promoting good health by contributing to access to medicines and vaccination. This could be achieved by deploying low-cost health care AI applications. This will help in establishing **accountability and non-discrimination practices.**

3.4.7.7 Issue 7: Social investment

Risks: Does the organisation invest its resources in initiatives and programs aimed at improving social aspects of community life like education, training, etc? Does it encourage community involvement in the design and implementation of projects?

Mitigations: An organisation should take into account the promotion of community development in planning social investment projects. It should consider partnering with other organisations, including government, business or NGOs. It should contribute to programs that provide access to food and other essential products. It will ensure practices of **sustainable development and reliability of the organisation.**

Questions addressing Community Involvement and Development:Does the product in any way discriminate between genders? Is the product availing employment creation? Is product responsible for developing new skills in people? Have all the stakeholders been identified? How are conflicts between stake communities resolved?

3.5 Mapping of Key Issues w.r.t other guidelines

KEY ISSUES	OECD	HLEG	IEEE	ASILOMAR	HEGENDORF	SDG	SWG3
No Poverty						■	
Zero Hunger						■	
Quailty Educatioin						■	
Clean Water & Sanitation						■	
Affordable & Clean Energy						■	
Decent Work & Economic Growth						■	
Industry, Innovation & Infrastructure						■	
Sustainable Cities & Communities	■		■	■		■	■
Responsible Consumption & Production	■		■	■		■	■
Climate Action						■	
Life Below Water						■	
Life On Land						■	
Peace, Justice & Strong Institutions						■	
Partnership For The Goals						■	
Human Agency & Oversight		■	■	■	■		
Technical Robustness , Safety & Cyber Security	■	■	■	■	■		■
Privacy Protection & Data Governance,Reliability		■	■	■	■		■
Transparency,Openness	■	■	■	■	■		■
Diversity,Non-Discrimination & Fairness	■	■	■	■	■		■
Environmental & Societal Well-Being	■	■	■	■	■		■
Accountability	■	■	■	■	■		■
Explainabilty & Interpretability			■	■	■		■
Solidarity, Inclusion, Social Cohesion		■	■	■	■		
Science-Policy Link			■	■	■		
Legislative Framework, Legal Status of AI systems	■		■		■		
Responsible/Intensified Research Funding				■	■		
Public Awareness, Education about AI & its rules			■		■		
Future of Employment			■		■		
Dual-Use Problem, Military, AI arms race			■		■		
Field-specific deliberations (Health,Military,Mobility)			■				
Human Autonomy	■	■		■	■		
Certification of AI Products			■		■		
Cultural Differnces in the Ethically aligned of AI sys.			■		■		
Protection of Whistleblowers		■	■		■		
Hidden Costs (Labeling, Clickwork,Energy,Resources)					■		

Figure 3.1: Mapping of Key Issues w.r.t OECD, HLEG, IEEE, Asilomar, Hegendorf, SDG, SWG3

This section proposes a mapping of various key issues w.r.t the guidelines from various international authoritative works (Other than ISO 26000). I have considered OECD [24], HLEG [9], IEEE [16], Asilomar [33], Hegendorf [14], SDG [3] and SWG3 guidelines for the

mapping.

The list of Key Issues in Figure 3.1 are extracted from the detailed documents of OECD (Organisation for Economic Co-operation and Development 2019),EU-HLEG(European Union-High Level Expert Group), IEEE's EAD (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019), Asilomar(2017), Hegendorf, SDG (Sustainable Development Goals by UN 2015), SWG3. This mapping was built in 2 parts:

1. **Guidelines Selection**: The selection criteria for OECD, HLEG, IEEE, Asilomar, SDG, SWG3 guidelines is based on the fact that the guidelines in consideration should not be older than 5 years and should be represented by a "government", "inter-government" or "multi-stakeholder" organisations. Another selection criteria was to keep only those documents in consideration which outline guidelines for Ethical and Trustworthy AI. Hegendorf is an exception in this case. Hegendorf[14] represents a detailed document which consists of key issues collected from various guidelines. It has been included here since the key issues reperesented in the document are compiled from 21 major ethical guidelines and represent almost major of the issues which need to be considered in AI.

2. **Key Issues Selection**: The key issues are extracted from the various guidelines mentioned above. There were multiple entries for the same issue under a similar name. For Example: Transparency and Openness in HLEG is synonymous to transparency in Hagendorf and synonymous to Failure Transparency & Judicial Transparency in Asilomar. The final key issue was therefore named Transparency and Openness. Good health from SDG is synonymous to Environmental and social well being in IEEE. So the final key issue was therefore named Environmental and social well being.Democratic values and diversity from OECD is synonymous to diversity, non-discrimination and fairness. So the final key issue was therefore named diversity, non-discrimination and fairness. In this way various synonomous terms were selected and are being represented by a common key issue.

This mapping 3.1 reveals that almost all the guidelines in consideration have addressed the key issues of **"accountability, Transparency, Non-discrimination, Diversity, Fairness, Technical Robustness, Privacy Protection, Environmental and Social-well being"**. It also

reveals that SDGs focus highly on sustainable development key issues which are less referenced in other documents(Note that the green colour in each cell highlights the positive occurrences of the issues).

3.6 Mapping of ISO 26000 w.r.t Key Issues

Figure 3.2: Mapping of ISO26000 w.r.t Key Issues (Part-1)

	No Poverty	Zero Hunger	Quality Education	Clean Water & Sanitation	Affordable & Clean Energy	Decent Work & Economic Growth	Industry, Innovation & Infrastructure	Sustainable Cities & Communities	Responsible Consumption & Production	Climate Action	Life Below Water	Life On Land	Peace, Justice, Strong Institutions	Partnership For The Goals	Human Agency & Oversight	Technical Robustness, Safety & Cyber Security	Privacy Protection,Data Governance,Reliability	Transparency,Openness	Diversity,Non-Discrimination & Fairness	Environmental & Societal Well-Being	Accountability	Explainability & Interpretability	Solidarity, Inclusion, Social Cohesion	Science-Policy Link	Legislative Framework, Legal Status of AI systems	Responsible/Intensified Research Funding	Public Awareness, Education about AI & its rules	Future of Employment	Dual-Use Problem, Military, AI arms race	Field-specific deliberations (Health,Military,Mobility)	Human Autonomy	Certification of AI Products	Cultural Differences in the Ethically aligned of AI sys.	Protection of Whistleblowers	Hidden Costs (Labeling, Clickwork,Energy,Resources)
4. Protection of the environment, biodiversity and restoration of natural habitats		✓		✓				✓	✓	✓	✓	✓								✓															
Fair operating procedures																																			
1. Anti-corruption													✓				✓																	✓	
2. Responsible political involvement																			✓		✓														
3. Fair competition																																			
5. Promoting social responsibility in the value chain						✓	✓		✓																										
6. Respect for property rights																	✓				✓														
Consumer Issues																																			
1. Fair marketing, factual and unbiased information and fair contractual practices																✓		✓			✓	✓													
2. Protecting consumers' health and safety									✓							✓				✓															
3. Sustainable consumption				✓				✓																											
4. Consumer service, support, and complaint and dispute resolution																					✓	✓													
5. Consumer data protection and privacy																	✓																		
6. Access to essential services																			✓				✓												
7. Education and awareness								✓	✓								✓						✓												

Figure 3.3: Mapping of ISO26000 w.r.t Key Issues (Part-2)

	No Poverty	Zero Hunger	Quality Education	Clean Water & Sanitation	Affordable & Clean Energy	Decent Work & Economic Growth	Industry, Innovation & Infrastructure	Sustainable Cities & Communities	Responsible Consumption & Production	Climate Action	Life Below Water	Life On Land	Peace, Justice, Strong Institutions	Partnership For The Goals	Human Agency & Oversight	Technical Robustness, Safety & Cyber Security	Privacy Protection,Data Governance,Reliability	Transparency,Openness	Diversity,Non-Discrimination & Fairness	Environmental & Societal Well-Being	Accountability	Explainability & Interpretability	Solidarity, Inclusion, Social Cohesion	Science-Policy Link	Legislative Framework, Legal Status of AI systems	Responsible/Intensified Research Funding	Public Awareness, Education about AI & its rules	Future of Employment	Dual-Use Problem, Military, AI arms race	Field-specific deliberations (Health,Military,Mobility)	Human Autonomy	Certification of AI Products	Cultural Differences in the Ethically aligned of AI sys.	Protection of Whistleblowers	Hidden Costs (Labeling, Clickwork,Energy,Resources)
Community involvement and development																																			
1. Community involvement				✓														✓			✓														
2. Education and culture			✓																																
3. Employment creation and skills development								✓	✓					✓							✓														
4. Technology development and access					✓						✓	✓		✓					✓		✓														
5. Wealth and income creation																		✓	✓	✓															
6. Health																	✓			✓	✓														
7. Social investment								✓																											

Figure 3.4: Mapping of ISO26000 w.r.t Key Issues (Part-3)

This section proposes a mapping of social responsibility issues of ISO 26000 w.r.t Key Issues derived in Figure 3.1. The reason for why these key issues are mapped to the specific social

responsibility issues, has been explained in Section 3.4.(They have been highlighted in bold towards the end of mitigation for each core issue)

The figures 3.2, 3.3 and 3.4 have been divided in three parts since each row and column consisted of more than 35 categories which was not being covered in a single readable page format. This mapping establishes that The Environment(Core Issue of ISO 26000) and its social responsibility issues all map to SDGs. Sustainability has been addressed extensively in ISO 26000. That is ISO 26000 aligns with SDGs in terms of ensuring sustainable development and practices. Most of the Human Rights, Fair Operating Procedures, Consumer Issues, Community involvement and development issues map with key issue of *accountability* and *transparency, security and privacy*. That is, it ensures to establish principles of social responsibility alongside with the main issues of AI. This mapping also serves as a proof that ISO 26000 covers the main key issues of AI and can be considered for establishment of concrete AI guidelines.

3.7 Mapping of ISO26000 w.r.t other guidelines

ISO 26000	OECD	HLEG	IEEE	Asilomar	SDGs	Hegendorf	SWG3
Human Rights							
1. Due diligence	✓	✓	✓				✓
2. Human rights risk situations	✓	✓	✓		✓		
3. Avoidance of complicity		✓				✓	
4. Resolving greivances	✓	✓				✓	✓
5. Discrimination and vulnerable groups	✓	✓	✓		✓	✓	✓
6. Civil and political rights	✓	✓	✓		✓	✓	✓
7. Economic, social and cultural rights		✓	✓	✓	✓		
8. Fundamental principles and rights to work					✓		
Labour practices							
1. Employment and employment relationships		✓			✓		✓
2. Conditions of work and social protection					✓		
3. Social dialogue		✓	✓				
4. Health and safety at work	✓	✓	✓		✓		
5. Human development and training in the workplace		✓			✓		
The environment							
1. Prevention of pollutions					✓	✓	✓
2. Sustainable resource use	✓	✓			✓	✓	✓
3. Climate change mitigation and adaptation		✓			✓	✓	✓
4. Protection of the environment, biodiversity and restoration of natural habitats		✓			✓	✓	
Fair operating procedures							
1. Anti-corruption				✓	✓		✓
2. Responsible political involvement		✓		✓	✓		✓
3. Fair competition							
5. Promoting social responsibility in the value chain		✓			✓		✓
6. Respect for property rights		✓					✓
Consumer Issues							

Figure 3.5: Mapping of ISO 26000 Social Responsibility issues w.r.t other guidelines (Part-1)

In this section, there is a mapping of international authoritative works that contribute to the establishment of trustworthy and ethical AI. This is an inspiration from the mapping established in paper written by David Lewis et.al. [20] which consists of a detailed comparison of ISO 26000 guidelines with IEEE EAD, OECD and HLEG based on normative statements.

Issue							
1. Fair marketing, factual and unbiased information and fair contractual practices	■	■					■
2. Protecting consumers' health and safety		■	■		■		■
3. Sustainable consumption					■		■
4. Consumer service, support, and complaintand dispute resolution		■	■				■
5. Consumer data protection and privacy	■	■	■	■		■	■
6. Access to essential services		■					■
7. Education and awareness					■		■
Community involvement and development							
1. Community involvement	■	■	■		■		
2. Education and culture					■		
3. Employment creation and skills development					■		■
4. Technology development and access		■	■	■			■
5. Wealth and income creation				■	■		■
6. Health		■					■
7. Social investment							

Figure 3.6: Mapping of ISO 26000 Social Responsibility issues w.r.t other guidelines (Part-2)

I propose to consider the mapping of social responsibility issues in ISO 26000 w.r.t guidelines mentioned in IEEE, HLEG, OECD, Asilomar, SDGs, SWG3 and Hegendorf. This mapping yields that most of the issues mentioned in other guidelines are covered extensively in ISO 26000. These documents are not directly comparable either in their purpose or intent. This mapping has been derived from the results and Figures 3.1, 3.2, 3.3 and 3.4 in Section 3.5 and 3.6.

It is observed in Figure 3.5 and Figure 3.6 that all of the considered guidelines have a number of statements which cover Human rights and Consumer issues. This is relevantly because all of these guidelines aim to establish mainly *accountability, transparency, data protection and privacy* of users. Also, Asilomar and Hegendorf do not address the environmental issues and have no coverage of Sustainable development issues. This is

because of the alignment of these guidelines mainly with the fundamental principles of AI. SDGs on the other hand not only address the environmental issues but also almost cover and map to all of the issues of ISO 26000. Asilomar, Hegendorf and SWG3 have almost no statements covering labor practices issues. This should be included since with the establishment of AI fears the displacement of labor, which is one of the greatest concerns which leads to statements like "AI will replace humans", "AI will be the end of humans".

Chapter 4

Technical Implementation

4.1 Introduction

This Chapter provides how the technical implementation of the AI Ethics Canvas tool. It states the importance of the tool and also proposes the development of AI Ethics Canvas as a semantic web application. It ends the discussion by stating how the unprecedented situations mentioned in 1.1 could have been avoided by the use of this tool

Section 4.2 aims to establish the arguments that need to be considered when designing a tool for assessing social responsibility issues. It also compares the performance of an existing tool named AI Blindspot which is designed for building Ethical AI application, and presents arguments w.r.t. this tool. This section is ended by considering that new tools should be devised for considering social responsibility in AI

Section 4.3 provides an overview of the Ethics Canvas. Ethics Canvas is considered as an inspiration for designing AI Ethics Canvas. It also states the perceived usefulness of The Ethics Canvas. It further states the previous work done on the Ethics Canvas design and proposes the present work.

Section 4.4 aims to propose the stages of completion of AI Ethics Canvas. It states a comparison done in the form of a table (4.1) between the Ethics Canvas Stages and AI Ethics Canvas Stages. Ethics Canvas stages of completion were considered as "building

blocks" for designing the stages of completion of AI Ethics Canvas

Section 4.5 provides the aims of the AI Ethics Canvas. In the first half, AI Ethics Canvas is proved in confirmation to the arguments proposed earlier in Section 4.2. In what follows next, a detailed explanation is provided of how can it be integrated in an Agile Workspace.

Section 4.6 describes the various features of the AI Ethics Canvas. It allows the user to edit, add, tag, comment collaboratively. Users can add other users as well. This also states that the AI ethics canvas is a successful tool which can be applied in workspace setting.

Section 4.7 aims to provide the importance of using the three canvas tools together. It further aims to provide an explanation for building AI Ethics Canvas, a semantic web application tool. It is done by proposing an ontology mapping of the various entities. It further explains how this mapping could be used to build a dictionary like approach for specific terms which will allow user to list synonymous terms together.

Section 4.8

4.2 Towards Artificial Intelligence Canvas

Google introduced Buzz, a social networking tool as a rival and answer to Facebook and Twitter. Google launched Buzz without adequate consideration of the ethical, societal, privacy or social responsibility impacts. Google automatically created instant social networks for users of its chat and G-mail services. The catch was that Google never asked the users whether they wanted a social networking account comprising of all people whom they have chatted with or G-mailed. As observed by a New York Times reporter,[35] 'E-mail, it turns out, can hold many secrets, from the names of personal physicians and illicit lovers to the identities of whistle-blowers and anti-government activists.' It resulted in a blast of criticism, which led Google to make changes to Buzz within a few weeks of its deployment. If it had carried out social responsibility impact assessment in advance of introducing Buzz, this firestorm might have been avoided.

Dr Barbara Ribeiro of Manchester Institute of Innovation Research claims that "The development of new Artificial Intelligence (AI) technology is often subject to bias, and the resulting systems can be discriminatory, meaning more should be done by policymakers to ensure its development is democratic and socially responsible." [28] There is a bridge between the policymakers who try to ensure trust in AI systems and the developers, designers who are involved in delivering the AI systems. Philip Brey in his paper about inculcating ethical and social responsibility impacts [8] in R&I argues that researchers, developers, designers are fundamental actors at a level where impacts in R&I are being explored because they have an understanding of technology(AI systems in this case) that many policy-devising people may lack. So there is a need to devise tools where R&I practitioners can include social responsibility issues and ethical impacts.

Considering the foregoing argument, it is proposed that the tool which aims to analyse social responsibility issues must be (1) Accessible to policy-devising people, non-ethicists or people who do not hold a background in addressing social issues. (2) Should allow people who belong to different backgrounds and roles to work together and identify the social issues. (3) The tool follows some existent guidelines which will help in ensuring the credibility of the tool and easier to convince R&I practitioners to abide by the tool. This will not only provide validation but in the times of arguments the existent guidelines will serve as a "handbook" or "referral text". Additionally the tool should facilitate an open-ended process of interpretation in a collaborative fashion to identify the potential AI ethical impacts and address the social responsibility issues arising.

There have been different tools designed for predicting unconscious biases or inequalities that an AI product can foresee. One such tool is AI Blindspot [4]. AI Blindspot was designed during Berkman Klein Centre and MIT Media Lab's 2019 Assembly program. Not only it is designed to be incorporated in a team's workflow and to provide oversights of issues that can generate harmful unintended consequences but has detailed User Interface. It consists of detailed cards where one can think in terms of questions proposed and predict certain risks. These cards can be downloaded in PDF format as well. It is hence not only a collaborative tool for non-ethicists, but it also addresses the upcoming shortcomings of an AI system. But AI Blindspot refers to no "established guidelines" or "legal text" which will help in ensuring the credibility of the tool. Not only that it does not address the Labour

Issues which is assumed to be the most harmful forthcoming of AI. Also there is no consideration of harmful effects on the environment which an AI product might foresee. This results in a need for a newer tool for analysing and addressing the shortcomings of AI systems and facilitate open-ended interpretations collaboratively also overcoming the shortcomings of existent tools.

4.3 Introduction to Ethics Canvas

Referring to a comprehensive literature review of R&I ethics approaches, Wessel Reijers recommends that "ethical technological design should be integrated into the day-to-day work of R&I practitioners; with guidance for discerning the ethical nature of technology design choices and how ethical principles can be balanced; different socio-technical alternatives considered; and stakeholder participation should be broadened while being guided by democratic principles" [27]. Ethics Canvas[31] is a collaborative tool designed for considering the ethical implications of a product/technology developed. The Ethics Canvas was inspired originally from a business implications modelling tool named The Business Model Canvas by Alexander Osterwalder of Strategyzer [30] and widely used for business development practice. Ethics Canvas consists of 9 thematic blocks structured together providing various factors to consider when listing down the ethical implications. The Ethics Canvas considers the fact that ethics is not simply about compliance and filling ethics clearance forms but requires an active reflection of issues during the design project. The Ethics canvas is currently available on the web [31] to use with or without an account. It can also be downloaded as a PDF (see Figure 4.1) or structure JSON of text in each block. It allows collaborative editing of an Ethics Canvas, shared comment threads on individual block entries and tagging of strings within a block entry. It is worth mentioning that the Ethics Canvas itself has followed an iterative approach over the last few years. In addition to the online website, David Lewis provides a handbook [32] that proves as a manual for the participants. It elaborates the meaning of different building blocks and the stages in which it can be completed. Nearly 500 students belonging to undergraduate, postgraduate and Ph.D. levels across computer engineering, science and business disciplines have used it till date.[21] Most of them apply the ethics canvas to an emerging technology, or business

projects they were working on or for the ethics evaluation of existing technical products. I considered the design of The Ethics Canvas as an inspiration for forming AI Ethics Canvas which will address social responsibility issues. Ethics Canvas addresses the ethical implications of a product, it needs to be redesigned for considering the social responsibility issues.

Figure 4.1: Ethics Canvas Design

I interned under Dave Lewis during June, 2018-Sept, 2018 where I worked on the next iteration of the canvas that intends to integrate The Ethics Canvas along with the Business Canvas Model. This approach was designed to allow users to investigate the ethical considerations of the technologies/ projects that the user is currently investigating business contexts of. In this way one can calculate both the profit and the ethical impact that a product may foresee. This integration will give a comprehensive view of how business and ethics can go hand in hand.

In this thesis, it is proposed to transform the ethics canvas into "Artificial Intelligence Ethics Canvas" and use it in parallel with The Business Model Canvas and The Ethics canvas. Adopting from my previous work which established an integrated model of business model canvas and the ethics canvas model, now AI Ethics Canvas will be added to the same. The AI Ethics Canvas has redesigned the ethic canvas to align it both with establishing social responsibility issues and their mapping into common ethical concerns of AI systems.

Note that one could possibly just integrate the business canvas model and AI ethics canvas model since AI Ethics Canvas has redesigned the ethics canvas to align it both with establish social responsibility issues and their mapping into common ethical concerns of AI systems. It will serve the same purpose. But here, I choose to integrate AI ethics Canvas Model alongside The Business Model Canvas and the Ethics Canvas so that one can consider all the aspects of the three tools. It will enable users to view what profit/money/business value a project/technology brings, what are the ethical implications of the project/technology, what are the social responsibility issues concerned with the project/technology.

4.4 Design of AI Ethics Canvas

The AI Ethics canvas 4.2 is a collaborative brainstorming tool that aims to identify social responsibility issues. AI Ethics canvas is a detailed 7 block structure. This 7 block structure of AI Ethics canvas is based on the 7 core subjects of ISO 26000 as block headings and the questions proposed in section 3.4 as the respective questions for the associated block heading. AI Ethics Canvas blocks can be grouped together and completed in 4 stages of completion. I have followed the approach of completion as in The Ethics Canvas [31]. For the AI Ethics Canvas, I revisited the "building blocks" of the Ethics Canvas(listed on the L.H.S of table [refer Table 4.1])[32] and proposed to complete the AI Ethics Canvas in the following stages(listed on the R.H.S of table [refer Table 4.1]): (On Page Number 63)

The Artificial Intelligence Canvas Model

Consumer Involvement and Development **1**	Consumer Issues **2**	Human Rights **4**	Fair Operating Procedures **6**	The Environment **5**
Does the product in any way discriminate between genders? Is the product availing employment creation? Is product responsible for developing new skills in people? Is the product a source of income creation or not? Have all stakeholders been identified and how will they be identified? Who will represent stakeholder communities? How are conflicts between stake communities resolved?	Can the AI System explain its decision to the affected consumers in a transparent way? Does the product in any way violate consumer's privacy by exposing his data? Is there enough explainability about the product as to why certain data is being asked for? Are essential services made available to the consumers about how will there data be used and disposed off?	How the product will affect basic human rights? Are there enough guidelines available for the product how it may be used, where it may be used? Does it violate any political, social and cultural rights of the stakeholders involved?	Could the product affect democratic decision making in an opaque manner? Are people contributing to data driving AI system in a fair and transparent manner? Does the product ensures fair competition amongst people using it or not using it? Does the product at present/later stage need political involvement? Does the usage of product amongst people results in corruption?	Is the product consuming resources by sacrificing the need of the future generation? Does the product displaces the activities by its user and stakeholders that incur higher or lower emissions? What happens with the product's use of resources, energy?
	Labor Practices 3 Could the product put people out of work? Does the product promote human development and training in work place? How the product could affect employment relationships? Does the product promote health and safety at work?		**Organisational Governance 7** Does the organisation owning the product take full responsibility of any damage reported to the society by its use? Does the organisation have made structured laws available which apply social responsibility?	

Figure 4.2: Artificial Intelligence Canvas Design

4.5 Aims of AI Ethics Canvas

AI Ethics Canvas is in accordance with the aforementioned 3 arguments in section 4.2. It is an easily accessible tool for the non-ethicists, developers, designers. It is a collaborative tool that actively reflects the issues of AI system during the entire design process and is inspired by the ISO 26000 which is an international organisation.

AI Ethics Canvas can be easily integrated into daily R&I practices. More and more companies are adopting Agile framework practices. [45] It is an iterative approach for developing and delivering products and features. Agile Framework follows the principle of Continuous Integration/Continuous Delivery (CI/CD). It means that the small features of the product are built over time, integrated with the product and are delivered to the user continuously. These practices iteratively analyse the needs and requirements of the customer and develops them over time to time. Similarly AI Ethics Canvas can be used iteratively and collaboratively. The R&I practitioners before delivering a product can analyse its social responsibility issues using AI Ethics Canvas. After addressing that issue via building a new feature or product, one can just comment or delete that issue from the Canvas and think about the new issues. This way not only a record will be maintained and

updated from time to time but it will be communicated amongst other teams and team members.

AI Ethics canvas exercise in a physical space consists of the Canvas sheet being printed out on paper, where the participants can write down their ideas. There is an online version of AI Ethics Canvas which is explained in Section 4.6.

4.6   Features of AI Ethics Canvas

AI Ethics Canvas is to be used alongside with Business Canvas Model and Ethics Canvas Model. The project was built using Vue.js framework for the front-end and Django framework for the back-end of the web application. The data is being stored in SQLite Database. The first phase of the project that is integrating the Business Canvas Model and the Ethics Canvas Model was developed in September,2018 where I worked on the front-end of the project. The second phase consists of integrating all the 3 models together. There were different features which are implemented for AI Ethics Canvas. The online features which are created for the AI Ethics canvas tool are:

1. On the first display the user is prompted with 4 options:

    - To Register/ Sign Up for using the tool. This allows the user to create his own account.

    - To Sign In for using the tool. If the user already has his account, he can log in with the credentials.

    - To Try AI Ethics Canvas Model, Business Model Canvas, Ethics Model Canvas Online. The User does not need to sign up for using this feature and the user can try the tool online. The data won't be saved.

    - To download AI Ethics Canvas Model, Business Model Canvas, Ethics Model Canvas in a PDF Format.

2. If the user signs up for the tool and logs in with the details, the next display is to create the project. A project can contain multiple canvases be it a business model

canvas or AI Ethics Canvas or Ethics Canvas. After creating the Project the user can select from the 3 canvases that are available.

3. It allows the user to add collaborators that is the user can choose who else can view in Project. This feature makes this tool a good option to be used in a technical or business environment. There is a button made available for adding multiple users. The user can also give administrative rights. If the user does so, the other user (promoted to admin) can add other collaborators. If the user feels threatened he can demote the other user and he won't be an admin anymore. Only the admins of the project will be allowed to add collaborators. Note that the user which will be added can only view the project in which he is added and not the other projects of the superuser.

4. In the Canvas one can add a Block Entry and add shared comment threads on the individual block entries. Users can also tag strings within a block and these tagged terms will be visible on the top of the canvas indicating them as important. Questions accompanying each block heading were added in the form of popups. These questions helped the user in thinking in context of the heading. These popups(Questions) feature was added for all the three canvases.

5. It allows multiple people to edit the same canvas at the same time. Whenever user say user1 is editing a canvas and the other user say user2 is viewing it, he can see when user1 is typing. It will be prompted on the screen saying "user1 is typing". This allows collaborative editing by multiple users. This communication feature was achieved by web-sockets. It is a computer communication protocol that establishes a persistent, bidirectional connection between a client and a server.

This application was hosted online and is available[1]. This website can be accessed by either TCD or SCSS network for now. The project is hosted online using tools named nginx and gunicorn. This figure 4.3 depicts the model that is being followed. The user types in the URL in his browser. Nginx cares about the requests from the world. Nginx catches this request and redirects it to the application. The application here is gunicorn. Gunicorn

---

[1]http://vma06.scss.tcd.ie

establishes a tunnel from nginx to Django app. Gunicorn serves the Django app content to nginx.
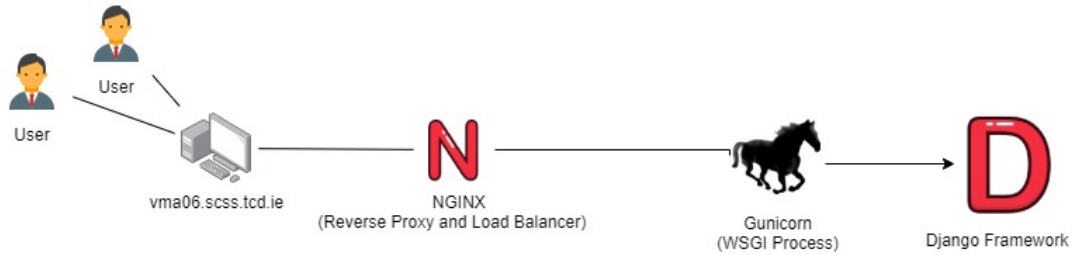

Figure 4.3: Structure for deployment of AI Ethics Canvas

4.7 Building a Semantic Web AI Ethics Canvas application

Integrating Business Canvas Model and Ethics Canvas Model with AI Ethics Canvas model will provide a multi-dimensional view of a product. Not only will it list out the profit a product can make but consider the impact of the product on the people and will help in ensuring the trust of people to use the product. Therefore linking the three canvases is an important part of our evolving approach of tools to support ethical and social responsibility issues. This can be achieved based on the structure of the 3 canvas models. All the 3 models are structured into a block, have individual ideas or segments populating each block based on the context. We explore how we implement this approach using semantic Web Technologies.

Semantic Web is an extension of World Wide Web through standards set by the World Wide Web Consortium(W3C)[42]. It is all about making computers understand the information so that they can generate better results. The fundamental principle of Semantic Web is to make Internet data in a machine readable format. Technologies such as RDF(Resource Description Framework), OWL(Web Ontology Language) are used for the same. RDF data model generates statements in the form of *Subject-predicate-Object* known as Triples. RDF can be thought in terms of other modelling approaches such as ER diagrams or class diagrams, the only difference is it models data in form of triples.

To support the broader use of the Artificial Intelligence Ethics Canvas, an open data

vocabulary is proposed to formalise and interlink the structured output of online canvas entries(which is available in JSON) using linked open data vocabularies. The vocabulary in Figure 4.4 provides support for the following features.
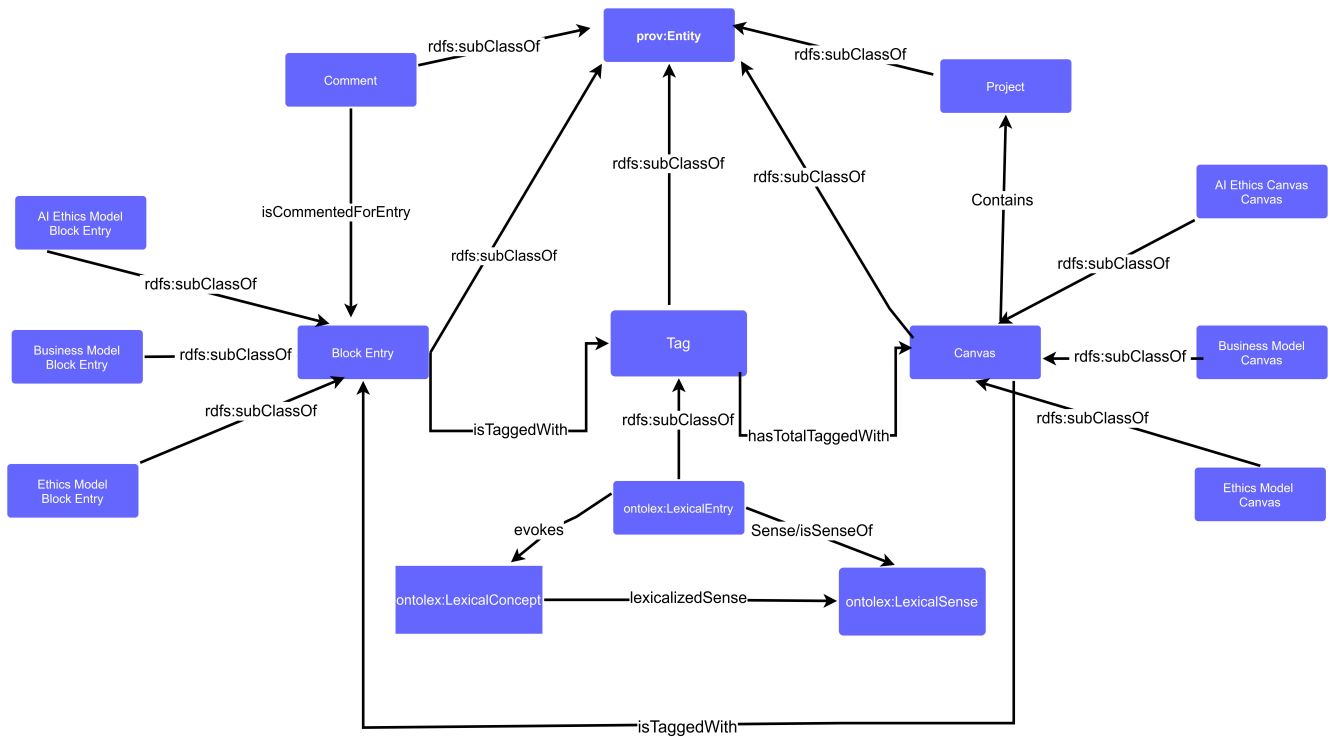

Figure 4.4: Vocabulary for integrating Semantic Web using Canvas Analysis

Grouped AI Ethics Canvas, Business and Ethics canvas perspectives: The vocabulary places all the 3 canvas models under the project class. This enables a multi-dimensional evaluation of a product using all of the 3 canvas types. This also enables us to contextualise the social responsibility issues analysis in the business and ethical context.

Canvas design evolution, innovation and variation: A general *Canvas* and *Block Entry* class is being used which is then subclassed by the AI ethics canvas types, Business canvas types and Ethics canvas types. The Block Entry types of the canvases are sub classed as well. However, In addition to Business Model Canvas there are many other canvas types that are focused on various other innovations e.g The GDPR canvas is emerging. This class design supports different canvas designs to be integrated, developed and used in the same project. The key principle here is to support the anticipated continuous evolution of

Artificial Intelligence Ethics Canvas.

Canvas Iteration and Logging of Entries: The fundamental principle of the canvas approach is its use to rapidly capture and modify ideas in an iterative, innovative environment. To support this, *Project, Canvas, Block Entry, Tag* and *Comment* classes are all declared sub-classes of *Entity* Class of the W3C Provenance Ontology (PROV-O). This allows a continuous tracking of the evolving canvas ideas across different canvases within the same project. The properties like *wasDerivedFrom, wasQuotedFrom* can be used to record and capture the history of the idea evolution and the actors involved *(prov:Agent)*. Also it can be successfully deployed to encourage broad participation and collaboration by different users. It will offer a log of ethical analyses of different entries for future reference in handling and down-streaming a social responsibility issue. It also allows to identify, analyse and react to the risk of the issues addressed.

Canvas and entry linking: Following from the current online canvas design of The Ethics Canvas, *Tags* can be associated with a canvas *block entry* and used for indexing entries to allow other users to search for the similar kind of tags within the projects which are available to them. This is useful in a public project or a private institutional projects. Therefore, I have listed *Tags* as a first-class object which allows them to be used on their own, separately from the block entries and canvases. So, this will allow a user to declare work involving a common issue that can be represented in the form of a *Tag* without the user having to publish the entire canvas and its block entries. This further enables the tag to be interlinked with other sources of knowledge such as existing ontologies.

Interlinking Tags and annotating Tags: The *Tags* listed in the canvases are proposed to be linked to an ontology which will be related to wider corpora of written material related to other social responsibility issues. As a proof of concept, *OntoLex* vocabulary was adopted for enabling them linking of a tag.[41] The OntolLex-Lemon vocabulary represents a vocabulary for publishing lexical data in ontology. An ontology is a way of showing the properties of the subject area and how they are related, by defining a set of concepts and categories that represent the subject. I have used OntolLex vocabulary specifically because it allows an ontology to be enriched with information about how the vocabulary elements being described in them are realized lexically, in particular in Natural Languages. The other ontology languages such as OWL and RDF(s) lack this enrichment of ontologies with

linguistic information, and hence is far from being able to capture necessary lexical information that Natural language application requires.

Tag in terms of Lexical Entry and Lexical Sense: Following Ontolex, A tag is represented as a *lexicalEntry*. W3C has defined a lexical Entry[40] as a unit of analysis of the lexicon that can have multiple forms and a base meaning associated with these forms. A lexical Sense represents the lexical meaning of a lexical entry when interpreted as referring to the ontology element in consideration. By using *lexicalSense* we can attach different properties to the *LexicalEntry*. Here, Lexical Sense is introduced to represent occurrences when the lexical entry is used with the given meaning. For example: Stakeholder and Contributor do not share a sense, even though they can be considered as synonyms. Similarly, "stakeholder is criticised" and "developed by stakeholders" are assumed to have the same lexical entry as they exhibit the same syntactic behaviours. Relevant terms are extracted a rough draft by ISO/IEC/JTC1 Subcommittee SC42 on AI. Most of the work in the paper is still in draft form, but it provides specific terms and concepts more specifically related to AI. In Figure 4.5 it is shown how different lexical entries like "stake-holders","stake-holders" are being annotated to the same Lexical Sense that is Stakeholder. These terms were then used to design an ontology. This implementation was achieved on the front-end of the AI ethics canvas tool as well.

In Figure 4.6 an "annotates To" button was added in the tag popup. The user can click on the button and see how the tag is related to others. In the Figure, user tags the term "stakeholders", this will annotate to "Stakeholder". So "Stakeholder" will be displayed on the interface. This implementation will be helpful in using tags to filter streams of related data for analysing the various issues which refer to the same term. This allows the user to see how the tagged term is being used in different canvases with the same context.

Tag in terms of Lexical Entry and Lexical Concept: A tag which is represented as a *Lexical Entry* can evoke multiple *Lexical Concepts*. A lexical concept consists of elements of meaning with different lexicalizations. For Example: Stakeholder and Contributor refer to the same ontology concept and as such they can be considered synonymous. Relevant terms are extracted a rough draft by ISO/IEC/JTC1 Subcommittee SC42 on AI which were useful in designing the ontology and referred to this relation. In Figure 4.7, It is shown how different terms extracted from the draft for risk issues were annotated as different Lexical
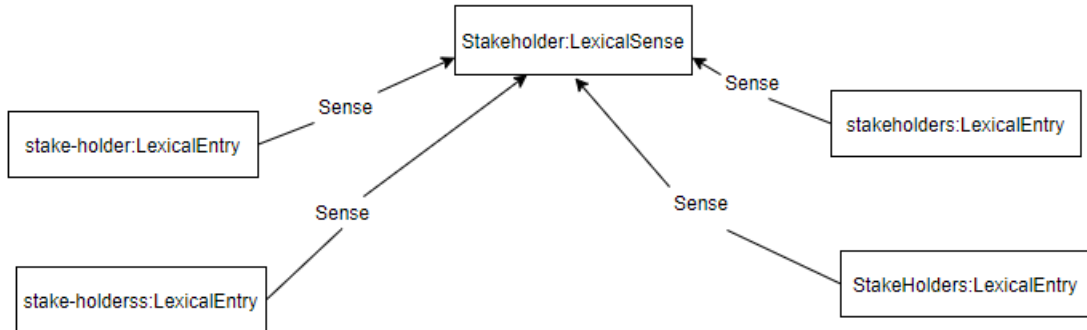
Figure 4.5: Different Lexical Entries being annotated to same Lexical Sense
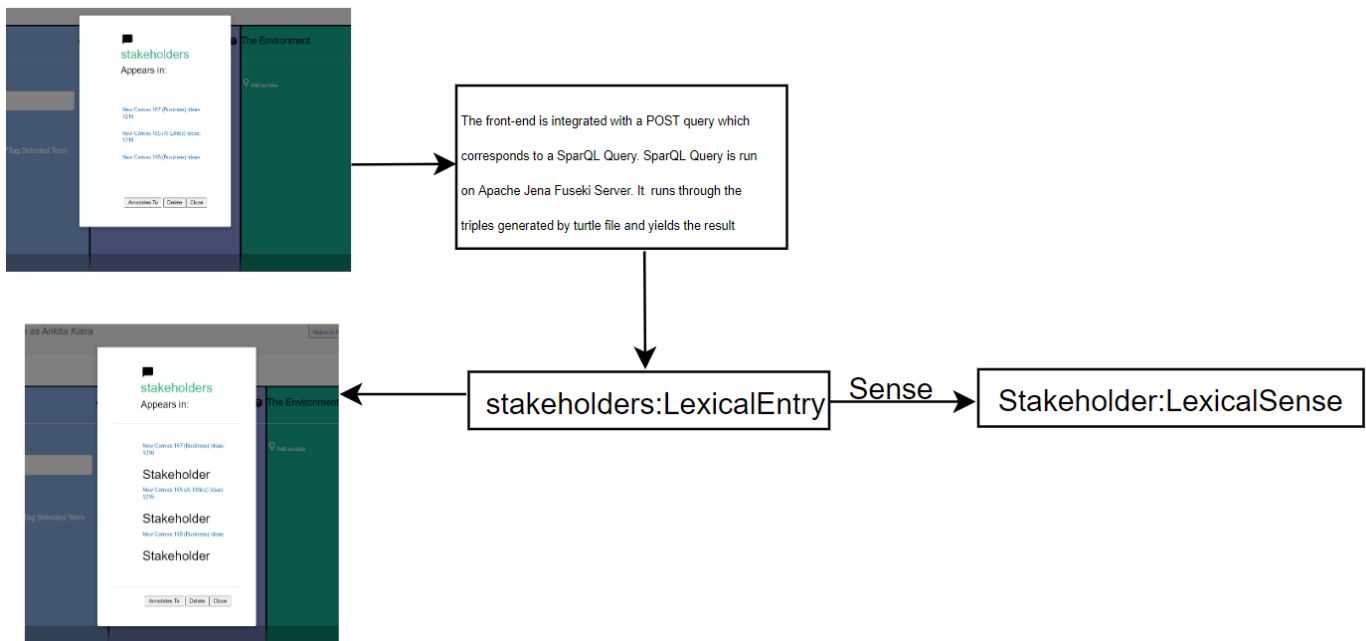

Figure 4.6: Annotating Lexical Sense for a Lexical Entry

Entries. Similarly In Figure 4.8, different Lexical Entries for Stakeholders are annotated to the same Lexical Concept.This implementation was achieved on the front-end of the AI ethics canvas tool as well.
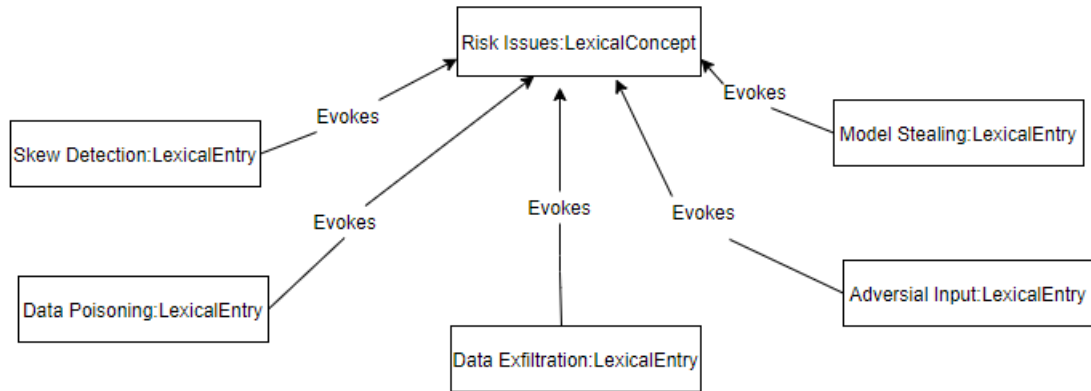
58

Figure 4.7: Different Lexical Entries for Risk Issues being annotates to same Lexical Concept

In Figure 4.9 an "annotates To" button was added in the tag popup. The user tags the term "Data Poisoning" this will annotate to "Risk Issues". So "Risk Issues" will be displayed on the interface as in Figure 4.9. This will help in filtering out superset of respective terms, thereby allowing users to relate these terms to each other. The user can certainly analyse the related data, which will allow him to think from a multi-dimensional viewpoint.

The tag annotations of the entries allows a better interlinking of similar terms even if they are being used in different forms. This establishes a "dictionary alike" approach and bridges the gap between different terms that are related to each other. This will benefit the user in numerous ways. Example: If user1 lists "skew detection", which will be annotated to "risk issues", and user2 lists "data exfiltration", which will be annotated to "risk issues" as well. User1 follows certain approach to mitigate the risk. User2 will be able to analyse how user1 achieved and can choose to make changes to mitigate his risk issue as well.

Figure 4.10 explains the flow of how the user is able to annotate a tag.

The Django comes with a built-in feature of export-import which allows the CSV files to be
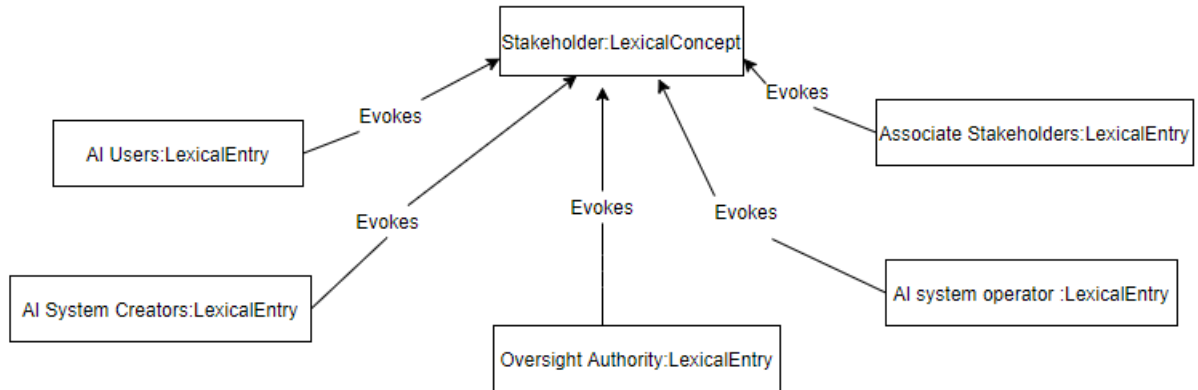
Figure 4.8: Different Lexical Entries for Stakeholder being annotates to same Lexical Concept

generated. CSVs for Project, Canvas, Tags, Comments and Block Entries were exported. These generated various datasets that could be queried. An ontology was built using tool protégé. These ontologies proved as a basis for designing the turtle files. Juma Editor was used to generate the turtle files. The CSVs for Project, Canvas, Tags, Annotations, Comments and Block Entries were uploaded in juma editor. The annotations CSV contains relevant terms extracted from the draft by ISO/IEC/JTC1 Subcommittee SC42 on AI. From the following, ontology generated in protégé, ontoLex vocabulary was integrated in the Juma Editor which generated the annotations to the tag. The turtle file was generated by the juma editor and uploaded on Apache Jena Fuseki server to run the SparQL queries. A SparQL query was designed which accepted the tagged term and looked for the Lexical Entry it mapped to. This SparQL query was then integrated into the front-end using the POST requests. In brief, whenever the Annotates To button was clicked the SparQL query was run on the Apache Jena Fuseki and generated the results which were displayed on the interface.
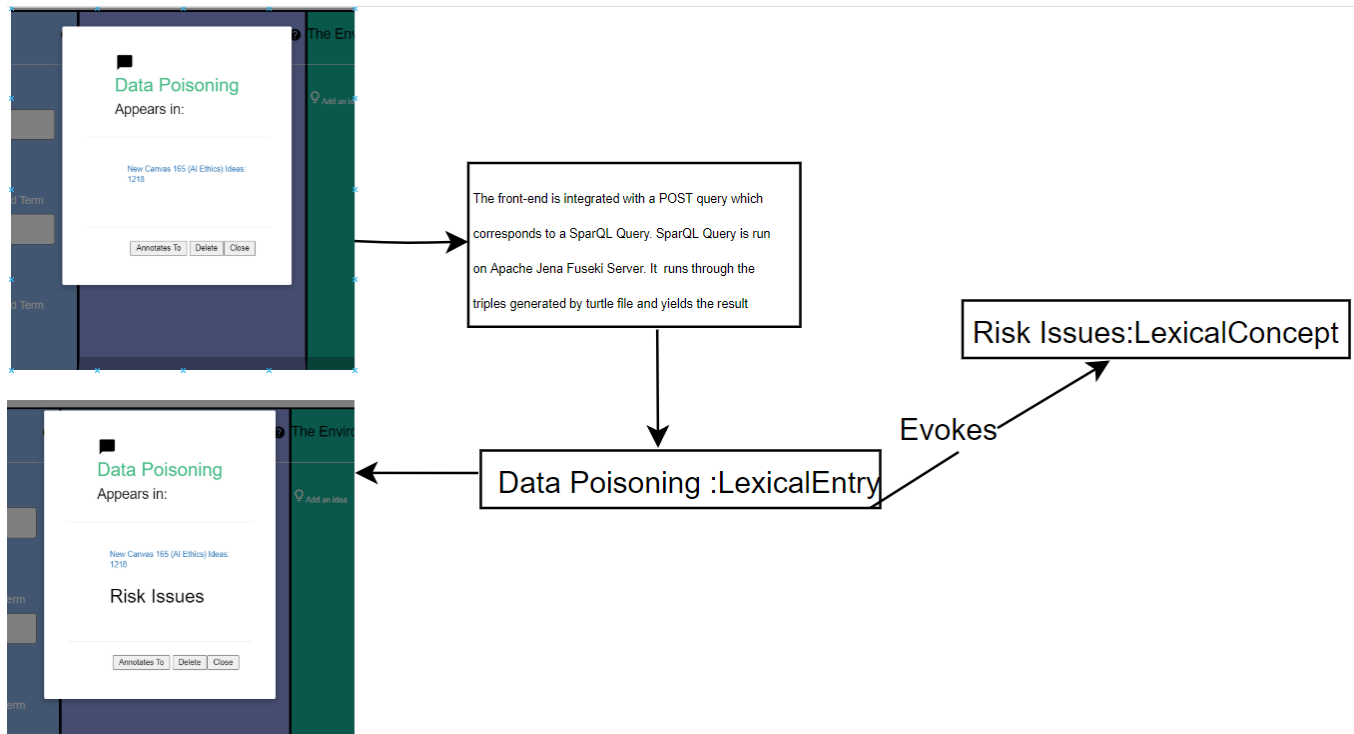
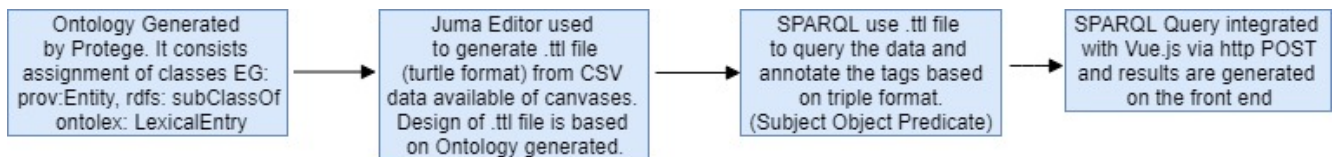Figure 4.9: Annotating Lexical Concept for a Lexical Entry

Figure 4.10: Structured Flow of printing Annotations

4.8 Summary

AI Ethics Canvas tool aims to address social responsibility issues by considering the core areas mentioned in ISO 26000 text. The controversies mentioned in Section 1.1 could have been avoided if such a tool would have been considered before. It allows the users to explicitly think of arguments in terms of specific categories. Example: If a user would have thought about the Fair operating procedures and provided a detailed explanation of how the organisation stores user data, the Cambridge Analytica controversy could have been avoided, or atleast predicted. Similarly, the controversy of Amazon(where it was claimed to discriminate amongst men and women) could have been avoided, if the issue of Human Rights and Community Involvement and Development was considered before. All in all, AI

Ethics Canvas can go a long way in solving such controversies.

Comparison of Stages of building Canvases	
Ethics Canvas	AI Ethics Canvas
First stage(blocks 1,2) identifies stakeholders involved that is the classes of individuals affected and types of groups affected	Stage 1: (Block 1: Consumer Involvement and Development) It revolves around practices of identification of stakeholders. Representing the stakeholder communities and resolving conflicts between stakeholder communities is the fundamental aim.
Second stage (blocks 3-6) identifies potential ethical impacts for identifies stakeholder. These impacts could be behavioral changes, conflicts which might arise, how are relations between individual and groups affected.	Stage 2: (Block 2,3,4:Consumer Issues, Labour Practices, Human Rights) It revolves around identification of social responsibility issues of AI system. These issues could be discrimination, complicity, resolving grievances which are covered under Human Rights. Or the issue of unemployment resulted by integration of AI technology which is covered under Labour Practices. Or the issue of privacy and security of data of users that is AI system should explains its decision to the affected consumers in transparent manner which is, covered under Consumer Issues.
Stage three(blocks 7-8) identifies the non-stakeholder specific impacts. It addresses the impact of system/service failure and resource use by innovative technology.	Stage 3(Block 5: The Environment) It addresses the impact on environment by the AI technology. It also addresses displacing activities of stakeholders which incur higher/lower emissions.
Stage four (block 9) consists of discussions around overcoming ethical impacts identified in previous stages.	Stage 4(Block 6,7:Fair Operating Procedures, Organisational Governance) It revolves around curbing the social responsibility issues by ensuring fair and transparent Operation procedures are adopted and making sure that organisation owns up to its responsibility.

63

Table 4.1: Comparison of Stages of building Canvases

Chapter 5

Evaluation

5.1   Introduction

This chapter aims to highlight the results of the survey that was conducted amongst students. This survey was conducted through questionnaire to assess the understanding of different categories of social responsibility issues. It also assessed the effort of the students when listing risks and mitigations for the social responsibility issues.

Section 5.2 aims to highlight the positive responses of Ethics Canvas and Business Canvas and suggest to assess the usefulness of AI Ethics Canvas

Section 5.3 assesses the facts and figures of the evaluation conducted amongst the students.

Section 5.4 highlights the comments of the students about the use of social responsibility categories for ethical assessment.

Section 5.5 concludes the discussion by welcoming more feedback on AI Ethics Canvas. It also suggests trying the AI Ethics Canvas amongst a vast number of users.

5.2   Overview

The Business Model Canvas is a tool that is accepted by all institutions and organisations. It has been positively assessed [22]. Also, usefulness of Ethics canvas as an ethical implications tool was assessed. [27] Similarly, In this research, the assessment for AI Ethics Canvas was carried. It highlights the usefulness of this tool to address the social responsibility issues. It is not possible to compare with other tools due to a lack of similar tools that can be used to address social responsibility issues. Therefore, The assessment for AI Ethics Canvas tool was assessed which evaluates its perceived usefulness amongst the users.

5.3   Assessing the usefulness and Evaluation of AI Ethics Canvas

An AI Ethics Canvas pilot was organised amongst students as a part of an assessment for the Module TUE00062: Engaging in Digital World, Today and Tomorrow. The students attended 2 lectures that familiarized them with the ethical and privacy issues that can arise when using AI applications. Lecture 1 was based on "Trustworthy AI for Digital Engagement", Lecture 2 was divided into 2 parts: "Regulation and governance challenges" and "Standardisation and Risk Mitigation". After this, 2 tutorials were held on Blackboard Collaborate Ultra where the students discussed about the stakeholders, ethical risks by social responsibility categories of a given AI application. These applications included YouTube Recommendations, Google+Fitbit+Next or Amazon Alexa. Students were split into four groups of 9 in breakout rooms for 20 minute discussion. One instructor was allotted to each group. In ideal situations, the website designed [1] would have been used when assessing the risks of social responsibility categories and collecting this feedback. Also, a feedback on the usability of the website would have been conducted highlighting more improvements in the AI Ethics Canvas tool.

All participating students were asked to fill in a survey which was circulated to them via an

---

[1]http://vma06.scss.tcd.ie

e-mail. This feedback survey was filled in by 10 students(lucky, because it simplified the analysis of the categories).

The survey followed a 5-point Likert scale, with a 1-point assessment indicating "very Easy" and a 5-point assessment indicating "Very Difficult". All the statements of the survey were formulated in the affirmative mode. The participants were asked about the (1). How easy/difficult did they find to identify risks for each of the social responsibility category? (2) How easy/difficult did they find to identify mitigations for risks for each of the social responsibility category? (3). Generalizing from your experience undertaking the ethics assignment, how easy/difficult did you find understanding each category as a principle for ethical assessment of a digital application? (4) Do you have any general observations about the use of social responsibility categories for the ethical assessment of digital applications? (5) Do you wish to be further contacted for trying out the online version of AI Ethics Canvas which identifies social responsibility issues? If yes, mention your email address. In what follows, the first three aspects(risks, mitigations, ethical assessment) are discussed based on reflections on the survey results.

The perceived usefulness of the social responsibility categories derived from ISO 26000 were evaluated extensively in the survey. The risks analysis (Figure 5.1) of human rights depicts that 50% believe that it is fairly easy to identify risks and 20% believed that it is difficult to think of risks in terms of human rights. For Labor Practices majority of students believe (70%) that it is an easy-moderate effort for analysis risks. Similarly, for the Environment,70% of students believed that it is fairly Very easy-moderate to analyse its risks. In Consumer Issues, 90% of students believed it is very easy-moderate to analyse risks. For Fair Operating Procedures, 60% of the students believed that it is moderate to analyse the risks, and for Community Involvement and Development, 60% which is again the majority believed it is Very-Easy-Moderate to analyse its risks.The Overall majority of the students believed that it was Very Easy to moderate to define Risks for each of the social responsibility categories. This suggests that AI Ethics canvas can be a useful tool to guide the participants in discussing risks that group members did not know or did not clearly think of beforehand. Hence AI Ethics canvas tool can be successfully used to establish socially responsible behavior.

The students were then asked to think of mitigations for the risks they have analysed in
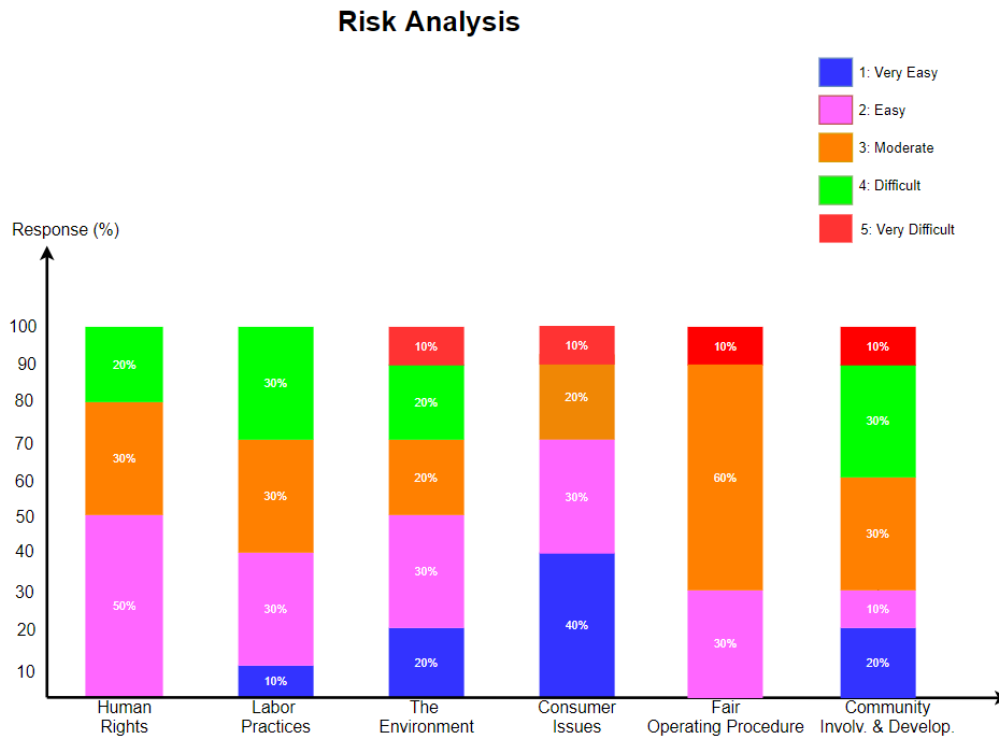
Figure 5.1: Evaluation of Risk Analysis of AI Ethics Canvas

each of the categories. The effort took for listing down the mitigations has been highlighted in Figure 5.2. The mitigations analysis of human rights depict that 80% believe that it is fairly easy to identify risks and 20% believed that it is difficult to think of mitigations in terms of human rights. For Labor Practices majority of students believe (60%) that it is a very easy-moderate effort for analysis mitigations. Similarly, for the Environment, 70% of students believed that it is very easy-moderate to analyse its mitigations. In Consumer Issues, 70% of students believed it is very easy-moderate to analyse mitigations. For Fair Operating Procedures, 60% of the students believed that it is moderate to analyse the mitigations, and for Community Involvement and Development, 80% which is again majority believed it is Very-Easy-Moderate to analyse its mitigations.The Overall majority of the students believed that it was Very Easy to moderate to define mitigations for each of the social responsibility categories. This suggests that AI Ethics canvas can be a useful tool to guide the participants into discussing mitigations of the risks analysed that group members did not know or did not clearly think of beforehand. These mitigations will help in preventing, predicting and foreseeing the risks and establish ethical behavior of AI. Hence AI Ethics canvas tool can be successfully used to establish socially responsible behavior.
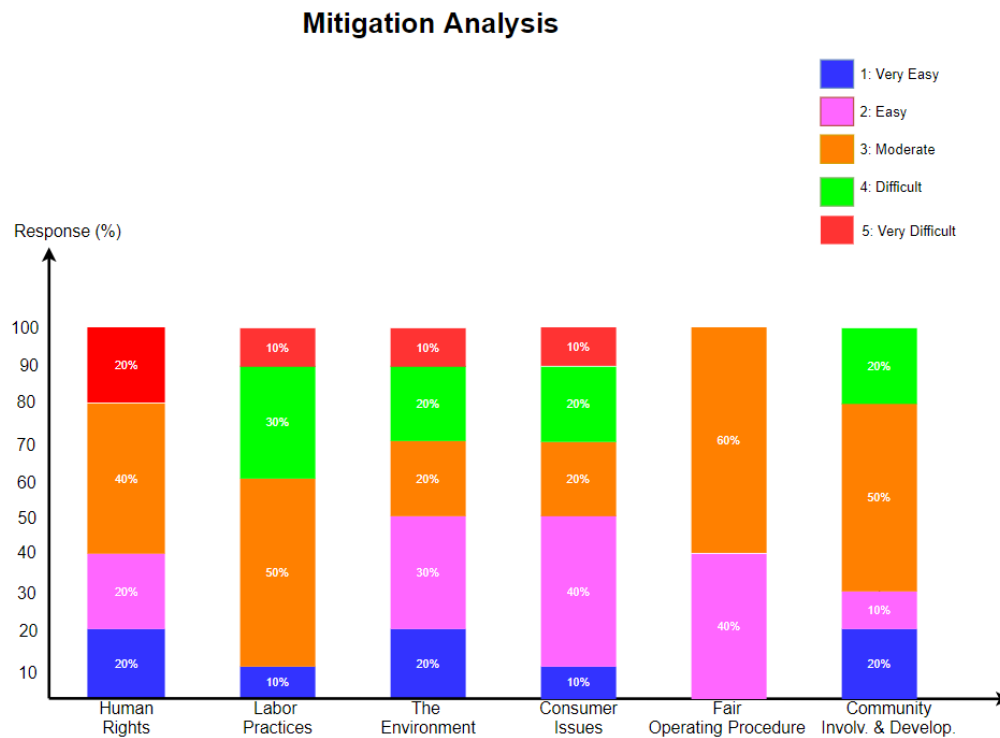
Figure 5.2: Evaluation of Mitigation Analysis of AI Ethics Canvas

The students were also asked how easy/difficult did they find understanding each category as a principle for ethical assessment of a digital application which will help in depicting importance of these categories when analysing the ethical implications of a product. The effort took for listing down the ethical analysis has been highlighted in Figure 5.3. The ethical analysis of human rights depicts that 80% believe that it is very easy-moderate to identify risks and 20% believed that it is difficult to think of ethics in terms of human rights. For Labor Practices majority of students believe (70%) that it is a very easy-moderate effort for analysis ethics. Similarly, for the Environment, 80% of students believed that it is very easy-moderate to analyse its ethics. In Consumer Issues, 90% of students believed it is very easy-moderate to analyse ethics. For Fair Operating Procedures, 80% of the students believed that it is very easy-moderate to analyse the ethics, and for Community Involvement and Development, 60% which is again the majority believed it is Very-Easy-Moderate to analyse ethics. Overall majority of the students believed that it was Very-Easy to moderate to define the ethical implications of a digital application in terms of the categories. This suggests that AI Ethics canvas can be a useful tool to guide the participants into discussing ethical implications that group members did not know or did

68

not clearly think of beforehand. These analyses will help to establish the ethical behavior of AI thereby ensuring the trustworthy nature of AI. Hence AI Ethics canvas tool can be successfully used to establish socially responsible behavior.
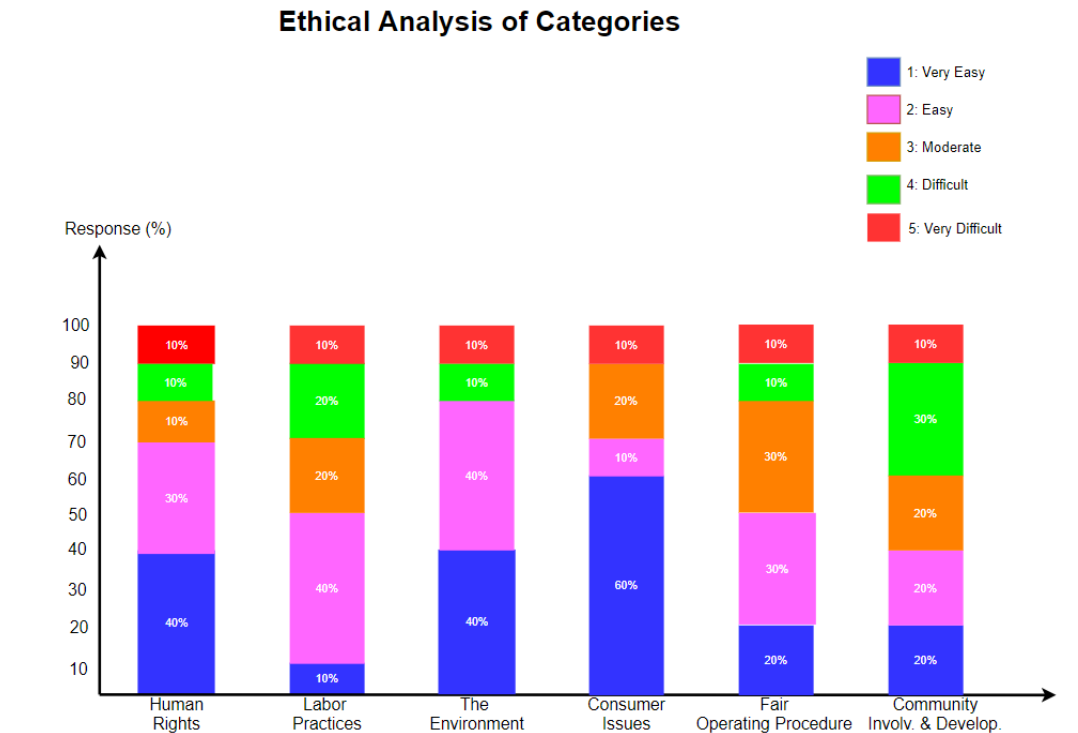

Figure 5.3: Evaluation of Ethical Analysis of AI Ethics Canvas

This analysis offers to establish that categories of Social responsibility which have been considered as the Block-Headings of the AI Ethics Canvas are easily understood by the participants. Most of the participants have found it moderately easy to list down the risks, mitigations of the AI application in the context of that categories. Most of the participants find it moderately easy to think of each category as a principle for ethical assessment of an AI application. Thereby, it can be concluded that AI Ethics Canvas tool can be successfully used when trying to establish socially responsible behavior of AI applications

5.4   Feedback on the AI Ethics Canvas

This section focuses on the Question " any general observations about the use of social responsibility categories for the ethical assessment of digital applications" in the

Questionnaire. 4 students wrote a response to this question which constitutes 40% of the overall participants(10). A student believed that "*I think more of a focus on how the categories apply differently to different stakeholders would have been helpful*". This could be considered in the questions accompanying each block category which will highlight the variances amongst stakeholders. Another student believed that "*almost all AI controlled applications have issues regarding human rights and "fair" operating procedure*". As the meaning of "fair" may differ from one person's opinion to another". This is true since the term "fair" is vast. It means ethical, socially responsible, just behavior and can be presumed differently by different people.

A student responded saying "*Most people would not be aware of the types of information gathered by companies and how the consumer becomes the product.It was very interesting to learn how your information is kept and gathered through AI*". This response establishes a valid understanding of the social responsibility categories. Yet another student believed "*I think if I had chosen another application/system, my answers would vary differently. Because my background is law and political science, my mind operates in the way that this assignment asked questions. It was not difficult for me to target the aspects that I have discussed more broadly in my politics modules. However, applying those broadly to the app I chose was difficult just because I did not conduct my own research, so relying on online resources was a bit difficult.*" This is a very generic case, since all the users belong to different backgrounds and their understanding of the categories are variable.

5.5   Summary

Overall, the results suggest that the AI Ethics Canvas is perceived as a useful tool to guide the participants when discussing ethical implications and social responsible issues as well as identification of relevant stakeholders. Moreover, the results also establish that the participant will consider revisiting their technological designs. Nevertheless, the results also indicate that the structure of AI Ethics Canvas will need to be improved to be more inclusive to list down risks and mitigations of the categories. Also this research is limited due to limited participation rate which might have led to biased results. The group of participating students could have coincided with the group of students that were more

positively engaged during this exercise. Hence, even though the former results establish AI Ethics canvas as a useful tool, further development and additional ways of assessing its usefulness will be helpful for future studies.

Chapter 6

Conclusion

This thesis aimed to establish the trustworthy and ethical nature of Artificial Intelligence Applications by addressing social responsibility issues. In the first chapter, various examples are listed in the motivation which serves as a need for tools and guidelines to ensure that AI operates in a fair and ethical manner. A research objective is framed which aims to narrow the gap between AI experts framing policies and technical community implementing AI by proposing tools which can be easily integrated in R&I sector. This research question is divided in two parts. First part focuses on proposing guidelines for AI (ISO 26000) and second part focuses on tools which are integrated with these guidelines and can be used (AI Ethics Canvas). The next Chapter "State of the Art" focuses on the need for Ethical AI and provides a definition of Artificial Intelligence which has been referenced from EU-HLEG guidelines. The contributions made by technical companies and various other organisation are highlighted in the next sections. This also proves that some of these guidelines are biased and can led to problems when applied to individual specific use cases.

The next Chapter Methodology gives and overview of ISO/IEC JTC 1/SC 42 and highlights the works of various Working groups in it. It also presents some concrete arguments which proof that ISO can be used when designing AI Guidelines which aim at establishing social responsibility. ISO 26000 has 37 core issues and 6 core areas which along with organisational governance can be used for proposing guidelines. Also, this chapter highlights the risks and mitigations each core issue of ISO 26000 will foresee. The AI Guidelines are presented in the form of mitigations. Towards the end of each core area,

there are various questions addressed which will be used in AI Ethics Canvas Sheet. These questions have been formed by conducting a detailed study of Risks and mitigations of each core issue and aim to establish social responsibility. Further, 3 mappings have been proposed. First Mapping represents some important key issues mapped with the guidelines (OECD, HLEG, IEEE, Asilomar, SDGs, Hegendorf, SWG3) which aim to explain the coverage of these issues in the latter. Second Mapping uses these key issues and maps them to core issues of ISO 26000. Third Mapping lists down a comparison between the various guidelines and ISO 26000 core issues. The aim of all these mappings is to serve as a proof that most of the key issues of AI have been addressed in ISO 26000. Also, most of the other guidelines address the same issues as addressed in ISO 26000.

Chapter 4 proposed the design of Artificial Intelligence Ethics Canvas which represents a tool that can be used by technical experts when ensuring socially responsible AI3 arguments have been proposed for the design of a tool which will be used for ensuring socially responsible AI. These 3 arguments were not satisfied by any other tool, that is why we turned to AI Ethics Canvas. This section further highlights how these 3 arguments are satisfied by the AI Ethics Canvas. . AI Ethics canvas is an iteration of Ethics Canvas tool which is being used widely. AI Ethics canvas uses the core areas from ISO 26000 as block headings and the questions formed in previous chapter as the block questions. Also the aims and features of AI ethics canvas have been proposed. A table has been designed which illustrates the stages of completion of AI Ethics canvas w.r.t the Ethics Canvas. Towards the end AI Ethics Canvas is designed as a semantic web application. An ontology has been proposed which highlights how the various aspects of AI Ethics Canvas are interlinked. Also Tags have been annotated to their Lexical Senses and Lexical Concepts. This section provides detailed explanation of how the semantic web features were achieved on the front-end of the AI Ethics Canvas. At the end, it provides an argument of how the controversies mentioned in Section 1.1 could have been avoided if AI Ethics Canvas tool was considered.

Chapter 5 represents the results of the survey conducted when AI Ethics Canvas was used in a classroom setting. The results yield a positive outcome. The risks and mitigations of categories of the AI ethics canvas were fairly easy for the participants to think of. This suggests that AI ethics Canvas can be used for designing mitigations which will help in

ensuring socially responsible nature of AI applications. Various feedbacks in the form of comments were also analysed in this Chapter. However, more studies will need to be done to further develop the Artificial Intelligence Ethics Canvas. All the assessments and the code is made available online. The github link to the code is [5]and the github code to the ontology is [7]. The link to the drive is [6]. It contains the survey, the video of the implementation, the presentation slides, the ethics Clearance documents(for conducting survey) and the Turtle file used to query.

Bibliography

[1] By Antoinette Price. Industry recognizes standardization will be essential to broad adoption of ai. https://iecetech.org/Technical-Committees/2018-03/First-International-Standards-committee-for-entire-AI-ecosystem#share.

[2] By Antoinette Price. Iso/ jtc 1/sc 42. https://jtc1info.org/sd_2-history_of_jtc1/jtc1-subcommittees/sc-42/.

[3] Vinuesa, R., Azizpour, H., Leite, I. et al. The role of artificial intelligence in achieving the sustainable development goals.nat commun 11. https://doi.org/10.1038/s41467-019-14108-y, 2020.

[4] AI Blindspot. http://aiblindspot.media.mit.edu/index.html.

[5] Ankita. The ai ethics canvas.

[6] Ankita. The link for the drive.

[7] Ankita. The ontology for ai ethics canvas.

[8] Philip Brey. Disclosive computer ethics. *SIGCAS Comput. Soc.*, 30(4):10–16, December 2000.

[9] European Commission. Ethics guidelines for trustworthy ai. https://www.academia.edu/38205904/ETHICS_GUIDELINES_FOR_TRUSTWORTHY_AI.

[10] European Commission's High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, 2019.

[11] Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. `https://dash.harvard.edu/handle/1/42160420`, 2020.

[12] Guardian. Alexa invading privacy- the dark side. `theguardian.com/technology/2019/oct/09/` `alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants`.

[13] Guardian. Cambridge analytica role. `https://www.theguardian.com/news/2018/mar/17/` `data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump`.

[14] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120, Feb 2020.

[15] IDC. "idc predicts ai contributions by 2021". `https://www.idc.com/getdoc.jsp?containerId=US43171317`.

[16] IEEE. Ieee ethically aligned design. `https://standards.ieee.org/content/dam/` `ieee-standards/standards/web/documents/other/ead_v2.pdf`.

[17] Information Week. Microsoft's satya nadella: 6 must-have ai design principles. `https://www.informationweek.com/cloud/` `microsofts-satya-nadella-6-must-have-ai-design-principles/d/d-id/` `1326109`, 2016.

[18] ISO contributors. Iso26000:2010 box 3, iso 26000 and small and medium-sized organizations (smos). `https://www.iso.org/standard/42546.html`, 2010.

[19] ISO Guidelines. `https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100418.pdf`.

[20] Dave Lewis, Linda Hogan, David Filip, and PJ Wall. Global Challenges in the Standardization of Ethics for Trustworthy AI, March 2020.

[21] Dave Lewis, Harshvardhan J. Pandit, Hannah Devinney, and Wessel Reijers. Towards an open data vocabulary for canvas driven innovation ethics. 2018.

[22] Garm Lucassen, Sjaak Brinkkemper, Slinger Jansen, and Eko Handoyo. Comparison of visual business modeling techniques for software companies. 06 2012.

[23] ntechz. Towards our artificial intelligence future.

[24] OECD:Better Policies for Better Life. Oecd ai principles overview. https://oecd.ai/ai-principles.

[25] Politico. "global ai battles ethics. https://www.politico.eu/article/europe-silver-bullet-global-ai-battle-ethics/.

[26] Pro Publica. Bias in criminal risk scores. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

[27] Wessel Reijers, David Wright, Philip Brey, Karsten Weber, Rowena Rodrigues, Declan O'Sullivan, and Bert Gordijn. *Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations.* Springer, 10 2018.

[28] Science Daily. Artificial intelligence needs to be socially responsible, says new policy report. https://www.sciencedaily.com/releases/2018/05/180510101249.htm.

[29] Global Solutions. What can the g20 do to implement ai principles and to shape global data governance?

[30] Strategyzer. Alexandar oxterwald. https://www.strategyzer.com/canvas/business-model-canvas.

[31] The Ethics Canvas. https://ethicscanvas.org/.

[32] The Ethics Canvas Handbook. https://ethicscanvas.org/download/handbook.pdf.

[33] The Future of Life Institute. Asilomar ai principles. https://futureoflife.org/ai-principles/.

[34] The New York Times. Article by the new york times, which analyses the penalty applied to google "under europe's data privacy law". https://nyti.ms/2HxCDNw.

[35] The New York Times. Critics say google invades privacy with new service. http://www.nytimes.com/2010/02/13/technology/internet/13google.html.

[36] The Verge. Amazon reportedly scraps recruitment tool which was biased.
https://www.theverge.com/2018/10/10/17958784/
ai-recruiting-tool-bias-amazon-report.

[37] The Verge. Google dissolves ai ethics board just one week after forming it.
https://www.theverge.com/2019/4/4/18296113/
google-ai-ethics-board-ends-controversy-kay-coles-james-heritage-foundation,
2019.

[38] The verges. The world first robot car's reported death.
https://www.theverge.com/2019/11/20/20973971/
uber-self-driving-car-crash-investigation-human-error-results.

[39] The wired. "in a huge breakthrough, google's ai beats a top player at the game of go".
https://www.wired.com/2016/01/
in-a-huge-breakthrough-googles-ai-beats-a-top-player-at-the-game-of-go/.

[40] W3C describing Lexical Entry.
https://www.w3.org/2016/05/ontolex/#LexicalEntry.

[41] W3C Lemon Ontolex Model. https://www.w3.org/2016/05/ontolex/.

[42] Wikipedia. https://en.wikipedia.org/wiki/Semantic_Web.

[43] Wikipedia. Three laws of robotics — wikipedia, the free encyclopedia, 2020.

[44] Wikipedia contributors. Iso 26000 — Wikipedia, the free encyclopedia.
https://en.wikipedia.org/w/index.php?title=ISO_26000&oldid=945906002,
2020. [Online; accessed 01-April-2020].

[45] Wikipedia on Agile Framework.
https://en.wikipedia.org/wiki/Agile_software_development.