DEVELOPING A MODEL FOR IMPROVING TRUST IN ARTIFICIAL INTELLIGENCE

ADITYA VASAN SRINIVASAN

AUGUST 26, 2019

This page is intentionally left blank

Developing a model for improving trust in artificial intelligence

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Management of Technology**

Faculty of Technology, Policy and Management

by

Aditya Vasan Srinivasan

Student number: 4755685

To be defended in public on August 26th 2019

Graduation committee

Chairperson	: Prof.dr.ir. M.F.W.H.A. Janssen, Section ICT
First Supervisor	: Prof.dr.ir. M.F.W.H.A. Janssen, Section ICT
Second Supervisor	: Dr. H. Asghari, Section Organization & Governance
External Supervisor	: MSc Micha Vink, Manager, PwC Accountants N.V.

This page is intentionally left blank

ACKNOWLEDGEMENTS

It is always a special feeling when we achieve certain milestones in our journey of life & they have to be cherished life long and it doesn't matter whether the milestone that we achieve is small or big. For me, this has been one of the major milestones in my life. I have to admit that I never thought that I would be doing my master's program in one of the best technical universities in the world and the beautiful moment started right from the time I received the admit letter from TU DELFT. But then seeing on the other side, I realized that doing masters is never going to be easy and especially for me, I anticipated that it would be really challenging to get back to study life after nearly been working in the IT sector for four years and the struggle was real as forecasted in the initial journey. However, I realized that the real value to the success and our commitment to the work will be hailed and respected only when we get past such challenges & hard times and this kept me motivated throughout the journey of my masters' life.

At this point, I would like to express my endless **gratitude** to the three most inspiring people who are indeed a part of my graduation committee. First and foremost, I would like to thank my main supervisor, Marijn Janssen for being a major pillar of support not just during my graduation project but throughout my master's journey program here in TU DELFT. Every word of encouragement & motivation from him meant a lot to me and I always feel a great sense of positivity after every meeting with him. His insights and valuable recommendations in each and every phase of my thesis journey have greatly influenced my graduation project. My second supervisor, Hadi Asghari for being a great mentor to me and I have to acknowledge that his insightful reflections and feedbacks have really assisted me a lot & gave me a clear direction throughout the entire phase of the graduation project. Finally, my coach from PwC – Micha Vink who always backs me up whenever I found challenges during the thesis journey and I have to be really thankful at this point of time for the immense assistance and mentorship that he has showered on me. I am extremely grateful for your endless support and investing your valuable time on my graduation project and more importantly, being a great motivator & an inspiring human to me.

I am extremely blessed & proud to work under the guidance of such inspiring professors and my coach from PwC – **Thank you** professors (Marijn Janssen, Hadi Asghari) and **thanks** to my coach (Micha).

Special mention of **credits** to these fabulous people from PwC – Mahdi Hazara, Michel Krishnadath, Miguel Brahim, Mona de Boer, Steven Hordijk, Aarav Vijay Sahu, Aakash Chavan Ravindranath, Ludy Rohling, etc. for offering a great deal of support during my thesis journey at PwC. Their timely inputs with regards to my thesis have to be really appreciated and respected. I would also like to take this moment to **thank** other colleagues of PwC who have contributed to my thesis (Interview respondents) & other well-wishers from the organization.

I would like to **thank** PwC from the bottom of my heart for giving me a wonderful opportunity to work on my graduation project. It is such an absolute delight to be a part of an amazing organization.

Every event of success, happiness, accomplishments that I have cherished in my life till date is because of eternal support, love, mentorships, and encouragement that I have been showered by several amazing souls and I personally feel that this would be the best moment to thank them as well.

To all my sisters - Devipriya, Monica, Priya, Rubiga, Jaseema, Karthika, Janani, Nahid, Nandhini, Balkis, Gayathri, Rahitha, Lavanya, Anusha, Jayu, Viji, Madhu, Sahana, Deepthi, Shreya, Sreenidhi, etc. for their constant support and eternal care showed on me. I have always been overwhelmed by your care & it means a lot to this brother. **Thank you & it means a lot to me.**

To all my school friends & in particular – Sai Krishna, Prashanth, Vignesh, Ranveer, Raghul, Sishir, Eka, Raghav Hari, Sandeep, Sharan, Bhuvan, Ramachandran, Pradeep, Sanjay, Bhaskar, Varun, Hari, Gokul, Gowtham, Ajay, Tarun, Nivedha, Deepthi, Sriba, Akshaya, Sai Janani, Bhavana, Shreya, etc. for always accompanying me whenever it was required. Our bondings are always special. **Thank you & it means a lot to me.**

To all my college friends & in particular - Anish, Alagu, Andrew, PandiThurai, Karthik, Anand, Anand Kumar, Guru, Chithra, Gayathri, Nisha, Lavanya, Sruthi, Mercy, Archana, Sameera, Amsarekha, Arjun, Indhu, Harini, Arief, Nishanth, Sriram, Ashwin, Naresh, Natraj, Linda, etc. for making four years of bachelors life an exciting journey. **Thank you & it means a lot to me.**

To my inspiring professors - Dr. Amrutha Radhakrishan, Mrs. Sumathi Poobal, Mr. Cloudin, Mr. Venkatesan, Mr. Rajaganapthy, Mrs. Mabel David, Mrs. Krupa George for showering constant support and finding out the best from me. Perhaps, all the major accolades and achievements that I have cherished during my bachelors' life is because of eternal care and motivation shown by Dr.Amrutha Radhakrishnan. **Thank you & it means a lot to me.**

To all my friends in Delft - Sownder, Prashanth, Anusha, Amrita, Santhosh, Sasidhar, Getssy, Saravanan, Prethvi, Yamuna, Prethivi, Krishna, Naveen, Praveen, Sharad, Balaji, Bala, Sidharth, Nivas, Sathish, Mohsen, MJ, Sanne, Ben, Joran, etc. for building so many beautiful memories and moments. It feels so good to realize that I have got such great friends who are always there to cheer and motivate each other and more importantly offering hands and care whenever it was needed. **Thank you & it means a lot to me.**

Moses, Kartik, Kalai, Hari, Anselm, Prasath, Prabha, Ashwin, Manoj, Pradeep, Ilango, Daniel, Jacob, Sabarish, Jones, Surya Rao, Surya Prakash, Jude, Heera, Ananth, Seetharaman, Srikanth, Srikanth Manda, Raviteja, Varun, Tamil, Maha, Raj, Krithika, Akshaya, Pon Varsha, etc. - Thank you for making every day of my work-life worthy and amazing. Let the journey of excitement and adventure continue for all of us. **Thank you & it means a lot to me.**

I have reserved the ending note to mention someone who is special to me and it is my family. All the good deeds and the success that I have seen till date is mainly because of my mother and my father. You have always been an inspiration and a role model for me. It is time for me to make your days exciting and create many more wonderful memories from now on. The most awaited message has now been delivered - Appa and Amma (Father and mother) - I did it !! **(Happy Tears)**

Thanks to my grandmother for bearing all the craziness that I did and showing me so much of love and care even now by enquiring about me through my mother. Paati (Grandma), you are always special to me.

I would also like to express my warm **greetings** to all my caring relatives, my amazing cousins and in particular, Arun for being so special in my life and being a great mentor.

*****Thank you TU DELFT *****

*****Thank you, God! *****

***** Vande Mataram*****

EXECUTIVE SUMMARY

We are at the frontline of the fourth industrial revolution where Artificial Intelligence (AI) is considered to be one of the biggest facets & highly transformative technology of this revolution. In simple terms, AI refers to the "**ability of a machine or a system that can perform human-like tasks**". The increasing availability of data, computing power & advances in the algorithms has really driven the development of AI in recent years. However, many industries & societies despite realizing the value of AI are still skeptical in accepting AI, especially when several controversial incidents have come into spotlights (For example, Amazons AI recruiting tool being biased to women, IBM Watson providing wrong recommendations, etc.) and this has increasingly raised the concern over the **trust in AI** & has become a major impediment while adopting AI.

Trust in general can be viewed as a set of beliefs that deals with benevolence, competence, integrity, & predictability and is important in all kinds of relationships (Mayer, Davis, & Schoorman, 1995). It can be seen in various contexts such as interpersonal trust in organizations, trust in virtual teams, trust in buyer-seller relationships, trust in automation, and now trust in AI. So, what does trust mean in the context of AI?

Using (Siau & Wang, 2018, p. 49) definition of trust in human-robot interaction, trust in the medium of AI can be defined as ***"the willingness of people to accept AI and believe in the suggestions, decisions made by the system, share tasks, contribute information and provide support to such technology".***

This kind of trust towards AI is already lacking as it is believed that people have already perceived in their mind that often, AI won't behave the way as intended, makes mistakes, provide harm to the society due to past issues posed by the technology. So, almost every stakeholders, potential users question upfront to the developers of the technology - Can AI explain the results?, How is AI using my data?, Is AI been governed? etc. and all these concerns address to one main question - **How can I trust AI or Whether AI can be trusted?.** Therefore, the problem statement can be defined as,

Although AI is about to transform the business model of every firms & industry and has the potential to provide benefits to the society & business, lack of trust from the clients, consumers & society as a whole is one of the major concern & challenge that is preventing the adoption of such technology.

Addressing the concerns posed by the clients & ensuring that AI models developed are trustworthy and responsible has now become one of the top priority & challenges for several technology-based companies. Scientific researchers, AI experts also agree to the fact that there is a need for improving the trust towards automation and AI in particular. Though several AI-based research institutions have already laid a few themes and principles as trust factors in order to trust AI, they don't contribute to greater value as they tend to generalize those themes as a whole and not clearly knowing on when such themes would come into the picture during the development. From the stands of scientific literature, there hasn't been substantial research done on the factors influencing the trust in AI despite the growing attention paid over the importance of trust in AI in recent times. At least, there is no enough study done on the concepts of trust in the field of AI from the management and socio-technical aspects and definitely, the scientific methodology to develop trusted AI models is already missing.

This research will focus mainly on improving the trust in AI and the main objective is to **develop a model to assess and improve trust in AI**. To improve the validity of the model & value of the overall research, the proposed model will be compared with some of the key themes & principles proposed by EU Commission, AI-based research institutions, and leading tech firms in order to ensure that the model has reflected those themes in this research. To realize a trusted AI model, factors influencing the trust in AI have to be identified. The major part of this research involves the identification of such trust factors that influence the trust in AI and see how these factors are being perceived by the actors involved in the development of AI in order to find the important trust factors from their perspective. The actors according to this study, would be **directors & partners, managers, data scientist, data engineer, AI experts & specialist, risk advisors & auditors** and these identified actors are the ones who are & who would be involved in the development of AI. However, it has to be agreed that more actors would be present when the level of automation gets complex and type of environment that the technology is targeting on. In this study, only the actors mentioned above would be considered in order to identify their perspectives of trust towards AI. So, the main research question can be defined as **What are the factors of trust that influence the trust in Artificial Intelligence from the perspective of actors involved in the development of AI?**

Trust can be seen in the broader context and in the case of AI, it can be determined based on human characteristics, environment characteristics, and technology characteristics. But, this research would focus only on the characteristics of the technology as trust factors. Additionally, at this point, it is important to recognize the value of data, as data is considered to be the main driver and input for AI and in such stance, it is critical to identify the trust factors not only towards the resulting AI model but also towards the incoming data that are fed to the AI models.

The research would mainly be employed with a qualitative study using an inductive approach in order to generate valuable theories as it is mainly supported by literature review, desktop research, interviews, and use of a case study. To be more precise, the research can be divided into two phases where the first phase involves the identification of potential factors that influences the trust in data and AI model and they are primarily derived from the extensive study done on the literature review & desktop research, and the second phase involves the identification of important trust factors from the perspective of actors involved in the development of AI.

To start with, literature studies were done, that initially gave an overview of the background of AI and the risks encircled around them. Secondly, the concept of trust, in general, was studied. Though there wasn't much research done towards the trust in AI, concepts of trust towards automation & its factors influencing the trust were studied extensively and were used as the main source of reference in identifying the potential factors towards the AI model. Furthermore, as the value of data is realized in this research, the dimensions of data quality (DQ) were studied and these dimensions would act as a trust factor of data. Based on the literature analysis, potential factors influencing the trust towards the data and model were identified. Interviewing the probable actors involved in AI development was considered as one of the data collection instruments and the main strategy. The main purpose of interviewing the actors is to understand their perceptions towards the trust in AI and identify the important trust factors relevant to the data and the model in the medium of AI. A total of **20 interviews** were orchestrated where **16 interviews** were conducted in the initial phase of the research and **4 additional interviews** were organized at the later stage of the research as a part of evaluating the initial version of the model.

Based on the analysis of the interviews, findings show that factors like **accuracy, auditability, bias-free, consistency, governance, explainability, privacy, reliability, security, transparency, and usefulness** were recognized to be the crucial factors as a whole. Upon segregating these factors clearly towards the data and the AI model, it was found that **accuracy, consistency, completeness, bias-free data, reliability** were identified to be the most important factors for trusting the data in terms of DQ. While factors like **governance** and **auditability** may not contribute to the dimensions of DQ but serve as core trust factor around the data process and lifecycle. Factors like **auditability, privacy, security, reliability, bias-free, transparency, explainability, governance, consistency**, etc. were highly regarded as important trust factors in the context of AI model.

Based on the actor's perspective towards their trust in AI, combined with the initial analysis of the literature, an initial version of a trusted AI model was developed. The model is formed of nine main phases. (1) Problem / Improvement Exploration, (2) Human-centered design, (3) Data acquisition, (4) Data preparation and validation, (5) Feature selection, (6) Model selection, (7) Model training and testing, (8) Model validation and deployment, and (9) Model monitoring. In each of this phase, trust factors that are crucial to be considered were identified and special attention has been paid from the third phase to eight phases as those were the identified phases where the value of data and model have to be assessed critically. Touchpoints were placed in the phase of human-centered design by identifying some of the key factors that need to be considered in that phase. Since the model developed is relatively new & more comprehensive in the first instance, it required evaluation with relevant actors to understand the feedbacks, insights over the model and making sure that those feedbacks are very well reflected before realizing the final version of trusted AI model. Furthermore, one of the cases was selected to evaluate the model and it was executed by interviewing the relevant actor (data scientist) involved in that case.

Based on the reflections and feedbacks, a final version of the trusted AI model was developed with six key prerequisites that were established upfront to realize the full value of the model. The model thus contains nine

main phases and in each of these phases, important trust factors with detailed indicators have been identified. The actors working on the respective phases have to ensure that the identified trust factors have been taken into account in order to place trust over the particular phase in the development of AI. On the other hand, the model would typically serve as a guide for the management levels like directors, and managers who could critically assess such factors using the indicators and see what kind of trust factor would require more attention in order to improve & establish a stronger trust over AI. Furthermore, the model as a whole can be shown upfront to the clients as that would certainly boost their confidence over the technology. Regardless of the above possibilities, the model would aid technology developers & management to provide a seal of trust to the clients, potential users and other stakeholders over the resulting AI products or solutions created. To realize the prime contribution to academic research, **reliability, consistency, completeness, bias-free, relevancy, and accuracy** were identified to be the essential trust factors of data. In the context of AI model, **explainability, interpretability, reliability, bias-free, usefulness, and transparency** were the essential trust factors. **Auditability** and **privacy & security under governance** were the core trust factors englobed around the data & model.

Several practical contributions and theoretical contribution have been established in this research. The main practical contribution of the research would be (1) trusted AI model itself, that would help the organization to assess the identified trust factors during the various phases involved in AI development, (2) identification of probable actors involved in the AI development & what kind of factors would those actors perceive in order to improve the trust in AI, etc. On the other side, some of the main theoretical contributions would be (1) identification of trust factors influencing the trust in AI model, (2) identification of the DQ dimensions in the medium of AI, (3) introducing new knowledge & insights to the scientific research about the trust concepts in AI. The research would finally end with the recommendations laid for future research and for PwC. Encouraging the scientific community and researchers to emphasize the importance of trust & its factors in the field of AI and mainly the value of data in AI, suggesting PwC to test the utility of the proposed model in the upcoming engagements and showing such models upfront to the clients for improving their trust towards AI, etc. were some of the important recommendations laid for future research and for PwC.

Keywords - Artificial Intelligence, Trust, Data, Data Quality, Models, Actors

TABLE OF CONTENTS

LIST OF FIGURES

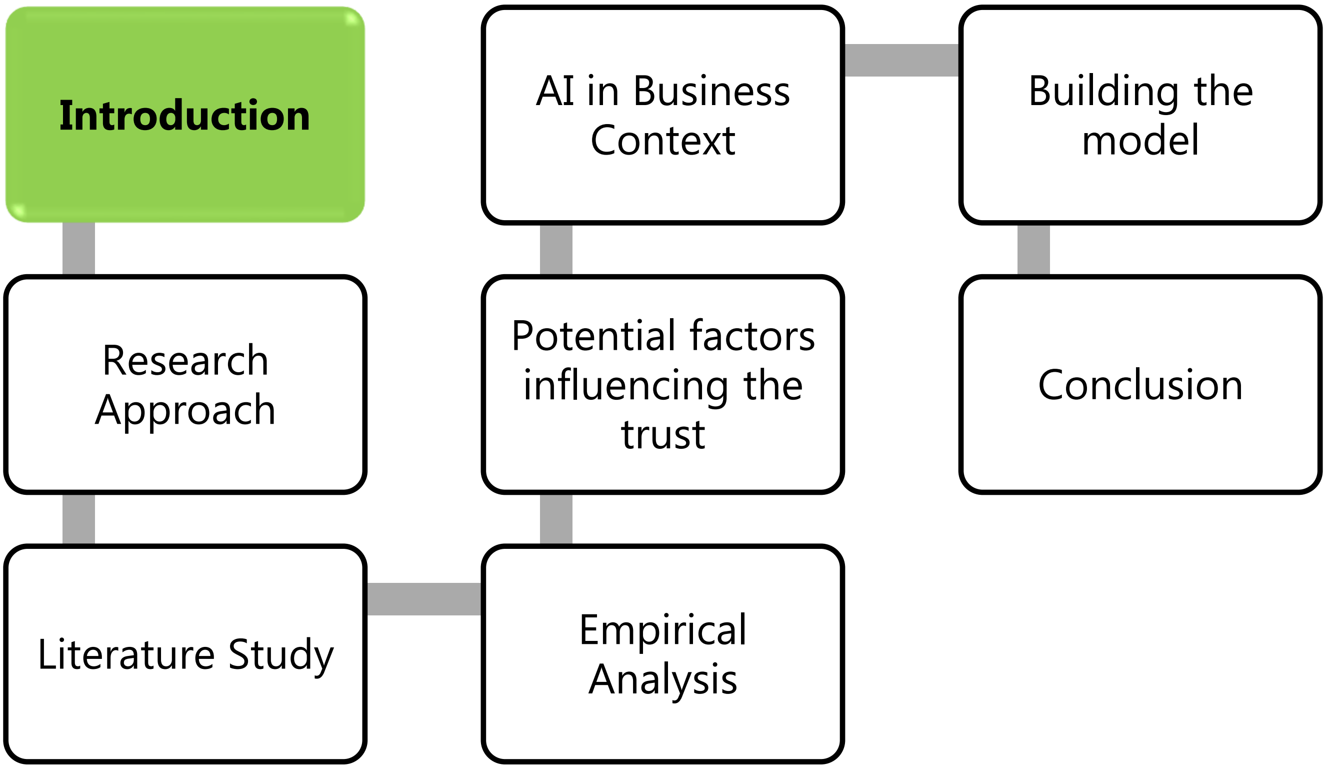
LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AGI	Artificial General Intelligence
ANI	Artificial Narrow Intelligence
ASI	Artificial Super Intelligence
ML	Machine Learning
NLP	Natural Language Processing
AR	Augmented Reality
VR	Virtual Reality
IoT	Internet of Things
DQ	Data Quality
IQ	Information Quality
PwC	PricewaterhouseCoopers
IS	Information Systems
ICT	Information Communication Technology
BPMN	Business Process Modeling Notation
MIS	Management Information Systems
EU	European Commission
SME	Small and Medium-sized Enterprises
MOT	Management of Technology

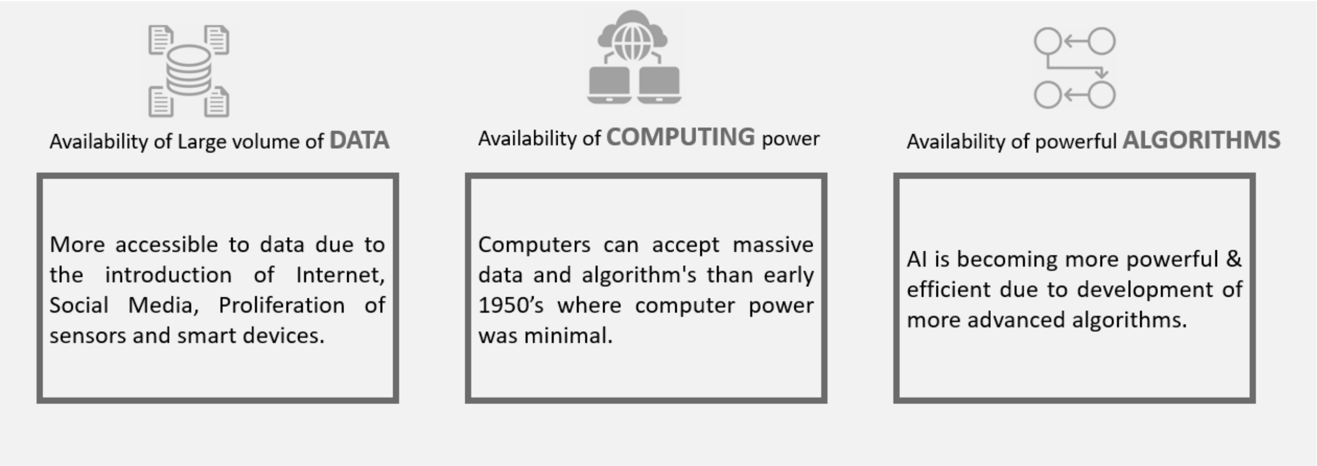
1

INTRODUCTION



1.1 BACKGROUND

In recent years, several organizations and industries are adopting digital transformations in order to open up their innovation challenges and opportunities to more eyeballs. New technologies such as Artificial Intelligence (AI), Blockchain, Augmented Reality (AR), Virtual Reality (VR), and the Internet of things (IoT) are shaping our world at a rapid pace and the possibilities are now beyond the imagination. They can completely define a new business model for several industries & organizations. There is no way that businesses and institutions can avoid digital transformation as they have to satisfy the increasing expectations of the customers and be competitive in the current market sector (PwC Advisory N.V., 2018). Perhaps, the most transformative technology that is available today is AI as it can ultimately transform every business in every industry (Marr, 2019). The impact that can be generated by such technology will be huge, loaded with plentiful benefits as AI has started to change everything right from automation to augmentation and beyond. **Availability of more data, increased computing power and advances in the algorithm** has driven the development of AI in recent times (IBM, 2018). It has started to become an important value for the companies as the technology could help them move closer to their customers in order to drive efficiency, enhance employee experience and capabilities and more importantly, decrease cost and increase revenues by automating the tasks (IBM, 2018).



Availability of Large volume of **DATA**

More accessible to data due to the introduction of Internet, Social Media, Proliferation of sensors and smart devices.

Availability of **COMPUTING** power

Computers can accept massive data and algorithm's than early 1950's where computer power was minimal.

Availability of powerful **ALGORITHMS**

AI is becoming more powerful & efficient due to development of more advanced algorithms.

Figure 1-Three main characteristics driving the development of AI

Through AI, Businesses can create intelligent products or services and can design a new intelligent business process in order to drive success (Marr, 2019). Some of the popular examples in which organizations are harnessing AI are AI-based digital voice assistants like Alexa and Siri, AI chatbots deployed as a part of customer support, facial recognition technology, and personal recommendation engine provided by platforms like Amazon and Netflix. As AI advances in the coming years, there is more space for greater value and capabilities that the technology can exhibit which can ideally provide enormous benefits for the economic growth, social development, and safety improvement (Wang & Siau, 2018). AI is not just about creating new industries and business models but in adding more value to the existing enterprises like (*2018 AI Predictions: Responsible AI: PwC*, 2018),

- Process of automating the complex & tedious process which is being done manually.
- Identifying patterns and trends in historical data.
- Strengthening human decisions by forwarding looking intelligence.

One has to clearly understand on the level of AI that is currently in place as AI can be classified into two types: Artificial Narrow Intelligence (ANI) or Weak AI, and Artificial General Intelligence (AGI) or Strong AI. ANI or Weak AI tends to focus on the repetitive, single and standard tasks and they are very powerful at routine jobs. Examples of Alexa or Siri that is existing today are the applications of ANI or Weak AI. As AI advances in coming years, it is expected that AI can exhibit human intelligence and can perform any cognitive task that a human can do which we refer them as AGI or Strong AI (IBM, 2018). The more the technology advances, the more it gets controversial and there is already a heated discussion from the perspective of academic researchers and experts as they are concerned that AGI can lead to Artificial Super Intelligence (ASI) (Müller & Bostrom, 2016). ASI are machines that can surpass human intelligence in all aspects. But, it was argued that ASI is just speculation up to date and having AGI could be the final outcome of the technology (Corea, 2018).

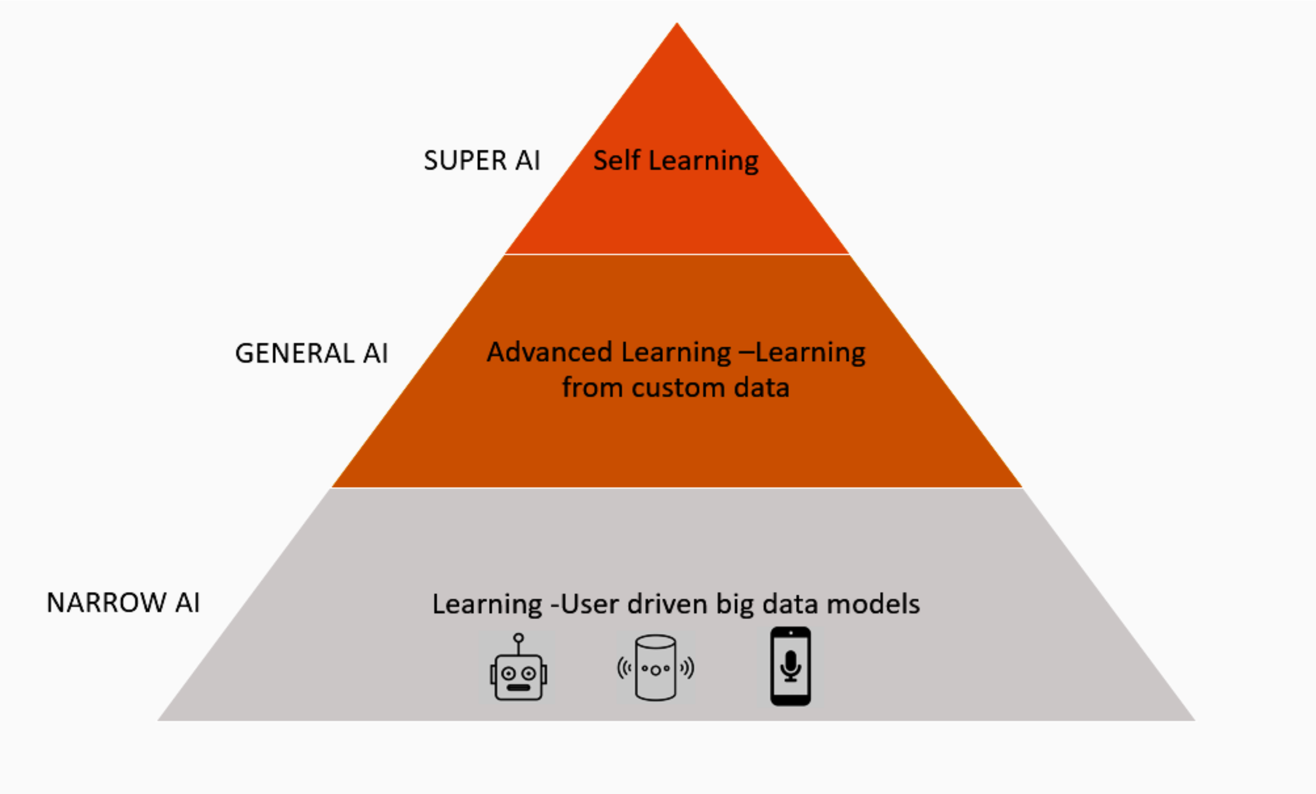

SUPER AI
Self Learning
GENERAL AI
Advanced Learning –Learning from custom data
NARROW AI
Learning -User driven big data models

Figure 2 - Classification of AI

Despite the value that AI can provide, they can also bring in an equal number of fears, risks, and challenges. Many industries realize the potential value of AI that can bring to their business, but they are reluctant to accept AI as there are multiple risks encircled around the technology especially when several controversial incidents came into the spotlights like Uber's self-driving car killing a pedestrian, Amazon's AI-based recruiting tool giving gender-biased outcomes, and IBM Watson providing wrong medical recommendations, etc. (Michals, 2019). The industries and potential users, therefore, expects the technology to be trustworthy & responsible. Several technology giants like Google, Facebook & Microsoft, AI experts and researchers are already emphasizing towards the importance of building a trustworthy & reliable AI. The technology advisory based companies like PwC, one of the big four consulting firms providing world-class assurance, tax, advisory services for the businesses would have a big task cut out when AI comes into the play as they would need to clarify the major concerns posed by the clients while doing audits on AI systems or providing AI-based solutions. Their concerns would be mainly on **"How they can trust AI", "Whether AI is safe & reliable",** etc. and addressing such concerns & building confidence over the client and the society with respect to AI would be the key priority and the challenge for organizations like PwC.

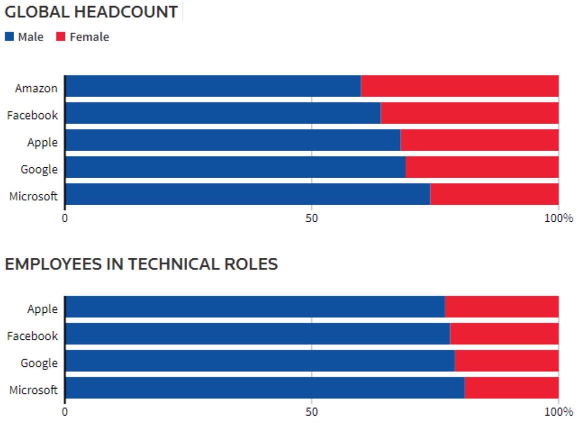
We would now see the actual problem and challenges in detail that are currently persisting and the problem statement that has been formulated in chapter 1.2 & 1.3. It will be followed by identifying the scientific gap in chapter 1.4. Once they are formulated, the main objective of the research as a part of the thesis work would be laid in chapter 1.5. Relevance to MOT study will then be formulated in chapter 1.6 and scope would be presented in chapter 1.7 to clearly define the boundary of this research. Finally, the outline of the thesis would be introduced in chapter 1.8 that will give a short snippet on the chapters to be followed in this research.

1.2 PROBLEM EXPLORATION

Although AI is marking itself as a promising and defining technology of our age by transforming the business model of organizations, providing possibilities & solution for consumers and society as a whole, its adoption would still be met with skepticism from various stakeholders as the technology is relatively new and untried. One of the major impediment in adopting such an AI-based model is how to trust the particular technology.

Imagine when two people meet for the first time, the first impression that would affect is the trust between these two people asking themselves **“Can I trust him?”**. Similarly, when it is seen in the context of AI, where the consumers or end-users experience AI for the first time, the first thing that would strike them is **“Whether I can trust this technology”** and this question on trust with respect to AI has become a major concern in recent times due to several issues that AI has created. People have already perceived in their mind that often, AI won't behave the way as intended, make mistakes, produces results unethically, show biases, provides harm to the society, etc. due to past issues posed by the technology (Burkhardt, 2019).

Research from MIT has stated that there are several issues which need to be addressed in order to trust AI in business and society, as one of the worrying concern is that people are not trusting decision, answers or the recommendations coming out from AI (Davenport, 2018). In one of the surveys of U.S. consumers, it was found that around 41.5% of respondents don’t trust any of the AI services like home assistants, financial planning, medical diagnosis, and hiring (Krogue, 2017). To give better clarity on how AI can pose a serious issue and lower the trust of consumers & society, we see one of the controversial incidents as mentioned in chapter 1.1 in detail here- ’’**The Amazon AI recruiting tool being biased against the women's**’’. The hiring team of amazon used AI to predict the job candidates on a score ranging from one to five stars. In the end, the team can hire the top 5 candidates based on the results that AI spits out. It was later found that the AI model was trained based on the historical resumes submitted over a 10-year period which showed the reflection of male dominance in their industry. As a result, it started discarding most of the women resumes (Dastin, 2018). This example has clearly shown a high risk of biases based on gender. Based on the below stats as in figure 3, the issue of gender bias would be a major risk when such technology giants like Google, Facebook, Microsoft, etc. are using AI tool in their recruitment. There would be an urgent need for serious re-evaluation if such AI models are used for classification, detection, and prediction based on race and gender (Crawford, West, & Whittaker, 2019).



Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.

Figure 3 - Stats showing the dominance of men over women in U.S. tech companies since 2017, copied from (Huang, 2018)

So, is this the only problem that is blocking the consumers or business leaders from trusting AI application? Certainly not, as this is one instance of several cases. The case of the autonomous car killing a pedestrian is calling for increasing concerns on safety, security, regulation, and accountability, while the case of IBM Watson providing the wrong recommendation to the patients, expresses concern over the lack of explainability over the decisions or the recommendations made (Michals, 2019). Such concerns towards the AI-based systems are already posing a significant risk not just to the users, developers but also to the governments & institutions (Wang & Siau, 2018). Though trust in AI is very closely related to the ethics, security, safety & privacy and other influencing factors with respect to the model, it also requires the trust in data in terms of data quality (DQ) in order to earn the user trust towards the technology as data is one of the huge asset to AI (*AI for Good Global Summit 2018 Report*, 2018).

Trust can't be built on one set of factor alone. If an AI model is fair but can't resist attack, it won't be trusted. If they are secured but can't explain the decision made, it won't be trusted. If they can explain the decision but the outcomes are bad, it won't be trusted. Trust in the context of AI is seen as a risk-based view. Mitigating the risks and challenges around AI would improve the trust in AI.

In the study of human-robot interaction, trust can be defined as the **"willingness of the people to accept robot-produced information and follow robots suggestions, share tasks, contribute information, and provide support to the robot"** (Siau & Wang, 2018, p. 49). Using this definition as a reference and with slight alterations, trust in the medium of AI can be defined as *"***the willingness of people to accept AI and believe in the suggestions, decisions made by the system, share tasks, contribute information and provide support to such technology***."*

Imagine, when an organization like PwC is about to provide AI-based solution to the clients in coming years, they would need to answer the major concerns posed by clients on AI which could be one of the top priority and challenges. It could be the client concerns and fear about the reliability, governance or it can be the users worrying about the biases. **For instance, a consumer retail company wants to anticipate the future customer demands in order to plan their logistics & supply chain accordingly and they approach PwC for a solution. PwC would use machine learning technique which is one of the subsets of AI, in order to learn from the orders that customers purchased earlier and provide the prediction accordingly. But what happens when the model makes a wrong prediction for reasons without justification or when the model misuses the customer information violating the privacy and policy aspects which could be the real risk. In such cases, users may not trust AI if they can't understand how it works and clients may not invest in AI if they can't see evidence of how it made its decision or when the technology lacks governance and compliance with respect to data protection**. If AI is put to work in other industry sectors like healthcare, transport, and financial services, the clients then would mainly expect the algorithm to be explainable on its decisions, whether there is a level of control over the technology and the safety concerns. It was also found in one other survey that most of the business leaders are reluctant to adopt AI and holding a step back as they are concerned about biases, lack of transparency, regulation & Control along with the stakeholder's trust as a major risk with respect to the technology (*2018 AI Predictions: Responsible AI: PwC*, 2018).

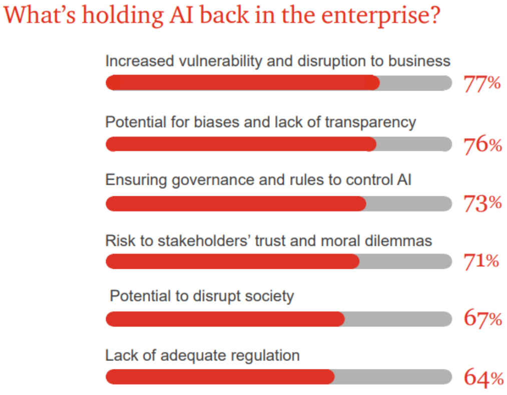


Figure 4 - Survey showing the list of issues surrounding AI adoption, copied from (2018 AI Predictions: Responsible AI: PwC, 2018)

While speaking to several technology enthusiasts at PwC, It was even realized that many clients are skeptical on how AI is being used in their case and the clients who approach PwC often question about boundary of the technology which can be seen as compliance, security & governance aspects and mainly the proof of trust with respect to potential technology outcomes and intentions. The same problem entitles when PwC does the audit for such technology as they have to provide assurance to the clients which could be a key challenge. Clients, therefore, expect a technology that is reliable and mainly the stamp of trust from the management of the organization and developers while implementing the technology and ensuring that proper controls and protocols are in place.

1.3 PROBLEM STATEMENT

In line with the above problem exploration, it can be inferred that AI has the ability to solve real-world & business problems at a wider spectrum with a game-changing impact but the lack of trust from the perspective of clients, end-users, and society towards the technology is certainly preventing them from adoption. Factors or trust dimensions with respect to data, AI model should be identified and analyzed in order to improve the trust of the technology. The resulting model incorporating such factors or dimensions should be developed with indicators to critically assess and improve the trust of the technology.

So, the problem statement for this thesis can be described as below:

Although AI is about to transform the business model of every firms & industry and has the potential to provide benefits to the society & business, lack of trust from the clients, consumers & society as a whole is one of the major concern & challenge that is preventing the adoption of such technology.

1.4 SCIENTIFIC GAP

Though one could find several academic papers that have been studied in the field of AI, they are inclined more towards the models of AI or the algorithm aspects that have been approached from a technical point of standards. In recent years, there have been growing concerns over trusting the technology as a whole especially from the management & socio-technical point of view as the technology is relatively new, untried, and has multiple challenges encircled around them especially with respect to trust. Researchers from multiple backgrounds have analyzed the position of trust mainly on the individuals, organizations, between the individuals & organizations (Lee & See, 2004). However, in recent years, researchers, AI experts, and technology giants also agree to the fact that there is a need for improving the trust towards automation, human-robot interaction, and AI in particular. After studying various literature and doing desktop research, It was realized that there hasn't been enough research done on the concept of trust in the medium of AI from management or socio-technical perspective till date despite the growing attention paid over the importance of trust in the field of AI. One could realize that enough research has been done towards the trust in automation, human-robot interaction and not AI in particular. Secondly, one could find several white papers, articles & report from several leading tech companies and AI-based research institutions like **AI Now** ("AI Now Institute," 2017)**, Partnership on AI** ("The Partnership on AI," 2018)**,** etc. recommending important themes and principles as a trust factor in order to build a trustworthy AI. However, they don't contribute to a greater value as they tend to generalize those themes as a whole and not clearly knowing on when such themes would come into the picture and where would such themes play an important role during the development. Additionally, there is no explicit representation of what kind of trust factors could possibly influence the technology especially from different perspectives of the actors involved in the development of AI as it is vital in building the solid level of trust by incorporating their opinions and concerns (*AI for Good Global Summit 2018 Report*, 2018).

The importance of data in the context of AI is highly valued in this research as data is the main fuel for AI. In this research, dimensions of data quality (DQ) would be considered as trust factors for trusting the data. There has been an extensive study done on DQ dimensions in certain specific sectors where data is considered crucial. Very recently, there were studies done on the dimensions of DQ in the context of big data but it was considerably scarce. Much surprisingly, dimensions of DQ in the medium of AI hasn't been studied widely so far despite realizing the fact that data is the key ingredient for AI and where the dimensions of DQ has a major role to play.

So, based on the problem exploration & statement, the scientific gap can be summarized as below:

- Not adequate research was done on the concept of trust & its factors in the medium of AI from management and socio-technical point of view.
- Not enough scientific study was done on the factors influencing the trust in AI from the actor's perspective involved in the development of Artificial Intelligence.
- The value & identification of DQ dimensions in the context of AI is missing in scientific research and this was realized after doing extensive literature study on DQ dimensions.
- There is no explicit differentiation on the trust factors towards the data, and the resulting AI model in the context of AI.
- Although many models and frameworks of trust are present in scientific literature, not many are related to the trust in AI, in particular.
- Scientific methodology to develop trusted AI model is missing.

1.5 RESEARCH OBJECTIVE

In order to provide a clear understanding of the problem and the challenges that are currently persisting, Problem exploration & statement were introduced in chapter 1.2 & 1.3. This leads the way to establish the main objective of the research. The development of AI solution is never easy as it would require careful planning, evaluation and assessment right from the identification of the problem to the monitoring of AI solutions post-deployment. In between these stages, comes the most important factor '**trust'** which needs to be initiated, established, nurtured and maintained. To establish the trust towards the technology, factors influencing the trust needs to be identified. Identifying and analyzing the trust factors/dimensions towards AI is certainly an essential requirement that we as a society and the potential users of the technology would expect before the adoption.

In partnership with PwC, research will be done to improve & strengthen the trust of the clients & consumers towards AI as it is important to have a trusted AI model that could drive the management of PwC to assess and improve the overall trust in AI. This can help PwC to provide a seal of trust to the investors and to the end-users. Also, the model can serve as a tool that could guide PwC to assess the technology (trust factors) as a part of auditing when it is being developed by an external organization and providing assurance to the clients. The model would contain detailed indicators with respect to each factor identified in every phase of the model and through this, the management can critically assess and determine what factors would still require attention in order to boost the trust of AI and overall the decision-making of the organization.

This research would emphasize only on trust aspects of the data and the AI model, which is expected to influence the overall trust in the technology. Other external trust factors will not be considered in this research.

So, the prime objective of this research would be to,

Develop a model to assess and improve the overall trust in Artificial Intelligence.

To achieve the main objective of the research, several sub-objectives have to be laid and answered. The below sub-objectives are presented in sequential order as follows,

- Identify the various risks involved in AI.
- Identify the probable actors involved in the development of AI-based solutions.
- Identify the DQ dimensions for the data and trust factor of the model/system from the literature study and desktop research.
- Identify the important trust factors from the perspective of relevant actors involved in the development of AI.
- Differentiate the trust factors that are specific to the data and the model.
- Prioritize and identify the main factors bounded to the data, model based on the initial analysis from the literature review, desktop research, and findings from the interview with the actors.

Additionally, the research would strive to achieve a couple of additional goals, as milestones apart from answering the main objective of the research so as to improve the validity of the proposed model and add value to the overall scientific rigor.

As the EU commission has recently published a document on ***ethical guidelines for trustworthy AI*** written by the high-level expert group on AI (European Commission, 2019), It will also be made sure that the resulting model is complying with the factors presented in the document as that could add high value to the model.

Many AI research institutions like ***Partnership on AI, Future of Life Institute, AI NOW Institute, AI for Good,*** etc. has been emerged to mitigate the possible risks & solve the pressing issues circled around AI. These institutions major themes include fighting biases, addressing the social implications of AI, promoting diversity and inclusions, making AI fair, explainable and accountable, etc. It needs to be ensured that the proposed model in this research has considered and addressed most of their themes that could possibly contribute new knowledge and insights to those institutions as well.

1.6 RELEVANCE TO MOT STUDY

The research is in accordance with the ***masters*** in ***Management of Technology*** (MOT) by using different concepts and methods that were learned during the academic year period, 2017-2019. The concept of the Business process studied as a part of Business process management & technology course has been introduced in the research as it would help to understand on how the process flow works in businesses when a new technology is introduced. Methods such as Business process modeling notation (BPMN) has been used to show the business process flow in a typical AI development. Strategies that were learned in this course especially in knowing the stakeholders can be laid as an initiating point to understand the values, expectation, and interest of the actors. As AI is one of the emerging and disrupting technology, most the knowledge acquired from courses like Emerging & Breakthrough technologies and Technology, Strategy & Entrepreneurship has been applied in this research. Insights and learnings acquired from Information Communication Technology (ICT) architecture & design, which was taken as one of the specializations in the master's program were used in this research. The value of data and their dimensions which could be seen here as influential factors of trust towards data has been inspired by that course. Finally, the methodologies and techniques for executing the research acquired from Research methods would aid in selecting the right research design & strategies, data collection techniques, sampling approach, and thereby making sure that the research complies with the scientific research rigor.

Motivation to carry out this research - AI has a great potential to solve the societal problems and business challenges if trained and implemented correctly. Humans and AI should evolve as a parental – child relationship, where parents nurture their child with good values, ethics & social norms and educate them by saying what is good and what is bad. Similarly, humans need to teach AI on

- How to do a task - by feeding the right data that is complete, accurate, reliable, etc. and has no form of biases (Gender, race, etc.) and choosing the right model and training them.
- More importantly, the social norms and values need to be taught to AI that can be accepted by society - By incorporating the social values, respecting & safeguarding the privacy of the users.

1.7 SCOPE

The research will focus on identifying the trust factors & developing a trusted AI model that could help PwC & other big firms in improving the trust in AI. Improving trust in AI can help the organization (PwC) to

- trust the technology by themselves (Management and Internal Stakeholders)
- providing a stamp of trust to the clients, stakeholders, and potential users of the technology.

The main intention of the research is to help organizations like PwC to be rather proactive by assessing the technology critically upfront by understanding what all factors would sound important at every phase involved in the AI development and whether it has been addressed in order to place the trust over the particular phase and proceed on to the next stage.

At first, to provide some shed light on AI, various dimensions of AI would be classified in the initial part of the research and one of the applications of AI, Machine Learning (ML) would be demonstrated on how they actually work. Nonetheless, the research won't cover the detailed or the technical aspects as they fall outside the scope of the thesis. Additionally, as the importance of data is valued in this study, only the dimensions of data quality (DQ) and information quality (IQ) in terms of processed data would be identified & considered as a factor in trusting data. However, the research won't cover or explain extensively on the strategies or mechanisms for improving the data quality. Additionally, it has to be noted that the dimensions used for DQ and IQ would be considered as same in this research and the purpose of saying IQ in this research is to convey that it's the quality of a processed data.

The research would be mainly supported with literature study, interviews and case study. In the first phase of the research, several respondents within PwC would be interviewed and it will be ensured that the respondents participating in this research fall in any of the below criteria,

1. The respondent is aware of the technology, its value & associated risks
2. The respondent has experience in working on AI and automation based projects.
3. The respondent has sufficient experience in working on data.
4. The respondent should either be a data scientist, data engineer, developer, AI expert, Partner or Directors at management level, Auditors, Advisors / Consultants.

In the medium of AI, trust can be determined based on human characteristics (competence, expertise, personality traits, understanding, etc.), environment characteristics (organization, workload, communication, etc.) and technology characteristics (reliability, transparency, usefulness, etc.). This research would mainly focus on the characteristics of technology as trust factors. More precisely, it is important to acknowledge at this point that the study mainly focuses on the concept of trust in terms of data and AI model only, other external factors of trust like **reputation, power,** etc. won't be considered in this research as well. Furthermore, It is important to note that this research would consider more of the objective attributes for placing trust in the technology and comparatively very less of subjective attributes. So, the objective attributes, for instance, would include factors like **explainability, reliability, usefulness, transparency**, etc. of the resulting AI model and factors like bias-free would be seen as a subjective attribute and it is considered as a trust factor in this research. Other subjective attributes like **loyalty, power, personal attachment, feelings,** etc. with respect to technology have been excluded from this research.

The literature on DQ and IQ can provide multiple dimensions that can potentially influence the trust towards the data. However, the research won't be considering all the factors or dimensions but only the most common ones found in the literature that is relevant to the context of AI would be identified and used in this research. Also, there would be a few dimensions that convey the same meaning. For instance, factors like **credibility** and **believability** refer the same though they are each considered as DQ dimensions and in such case, only one of the dimensions would be considered. Some of the most frequently mentioned dimensions are accuracy, completeness, consistency, accessibility, relevancy, reliability, etc. and this kind of dimensions would be used in this research. Factors relevant to the AI model, in particular, is considered scarce in the literature. So, factors relevant to automation, robotics and system quality, in general, would be used as reference but the research would consider only the factors that make sense to the AI model. Factors like reputation, process with respect to the automation would fall outside the thesis scope.

Though the final model takes into account on most of the major risks like security, data privacy, data biases, lack of governance and explainability, etc. encircled around AI. It is also believed that risks like automation taking over human jobs, AI taking control over humans are also considered important but those risks won't be addressed in this research as the main intention of the research is to improve the trust in AI so that the technology can be accepted by the potential users of the technology, clients, and stakeholders.

1.8 OUTLINE OF THE THESIS

The thesis outline will be presented as follows,

Chapter 2- This chapter presents the research methodology up front followed by the research framework for this study where the schematic representation of the research framework would be portrayed. The chapter formulates the main and sub-research questions that will help to achieve the main objective of the research. Additionally, the chapter lays out the guiding principles and recommendations for qualitative research in IS which would be followed in this research. Finally, an overview of the data collection methods and analysis would be presented.

Chapter 3- This chapter provides an extensive analysis of the theoretical concepts and it follows a sequential pattern. First, an overview of AI and the risk encircled around AI would be presented. Second, the concept of trust, in general, would be studied and this navigates to the study on trust in automation where various factors of trust and models would be discussed. Realizing the value of data in AI, dimensions of DQ would be studied in depth in order to determine the dimensions of DQ in the context of AI. Finally, the core themes or principles laid by several AI research institutions and tech companies would be identified.

Chapter 4 – This chapter provides an overview of AI in the business context. The chapter would start with the identification of stakeholders in the AI development followed by the overview of the business process which would be presented using BPMN. The chapter also elaborates on how AI works and provides a demonstration using a sample use case. Finally, the chapter would end by emphasizing the importance of trust by identifying the trust issues in various phases of the development of AI.

Chapter 5 – This chapter identifies potential factors that influence the trust in the data and AI model based on the in-depth analysis from the literature review and desktop research. Based on the identification of the potential factors, the working hypothesis would be formulated for each factor & a conceptual model would be developed accordingly which would be validated at the end of the research.

Chapter 6 – This chapter first elaborates on the interview process conducted with the actors and lists some of the interview techniques and process that was followed while conducting the interview. This would be followed by the methods used for analyzing the interview transcripts. One of the key aspects of this chapter is the analysis and findings from the interview which would be elaborated in detail in the final section of this chapter.

Chapter 7 – This chapter first identifies the main phases of the AI development which is the main requirement for building the model and in each phase, the relevant factors would be identified with detailed indicators that are derived based on the combined analysis of the literature review, desktop research and initial findings from the interview. The initial version of the model proposed would then be evaluated using interviews and a case study. Based on the reflections and feedbacks, a final version of the trusted AI model would be realized. To improve the value of the model and overall validity of the research, the proposed model would be compared with some of the core themes and principles laid by EU commission, leading tech companies, and research institutions

Chapter 8 – This chapter would first discuss the answers to the main & sub research questions that were formulated in the early phase of the research. Second, the essential factors of data and the AI model would be determined in order to realize the primary contribution to scientific research. The chapter would also discuss the model reflection and this would be followed by validation of the working hypothesis and conceptual model. The chapter would then elaborate on the contributions made in terms of theoretical and practical contribution followed by the limitations encountered in this research. Finally, the chapter would end with the recommendations for future research and for PwC.

1.9 READERS GUIDE

In order to avoid confusion to the readers, several terms used in this report would interpret the same meaning and these terms are presented as below:

ACTUAL WORDS	RELEVANT WORDS
AI model	AI solution, AI systems, AI applications
Technology Investors	Client
Factors (in terms of the trust)	Dimensions (in terms of the trust)
Consumers	End-users (in the context of this research, Consumers, and end-users imply the same)
Participants	Respondents, Interviewee

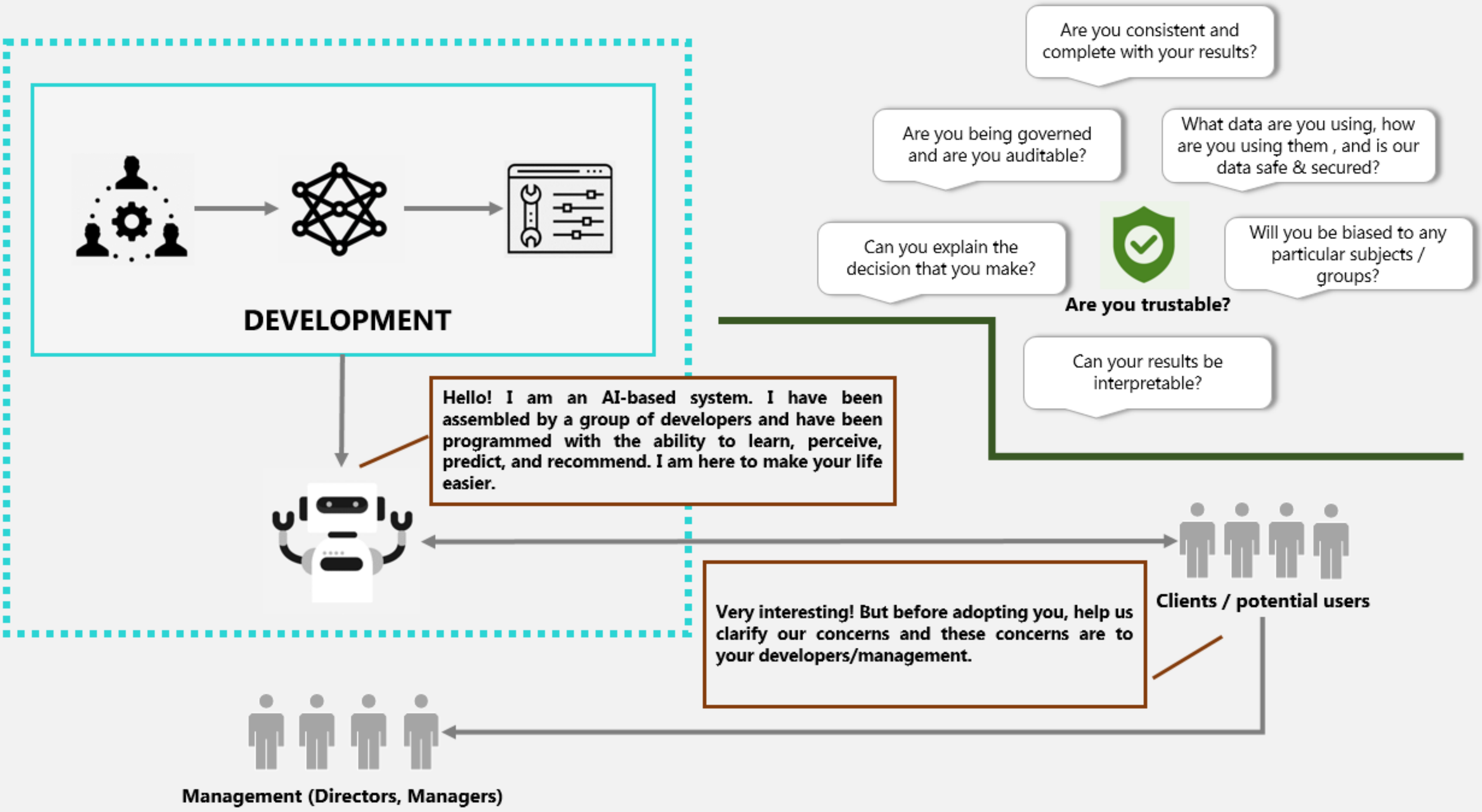
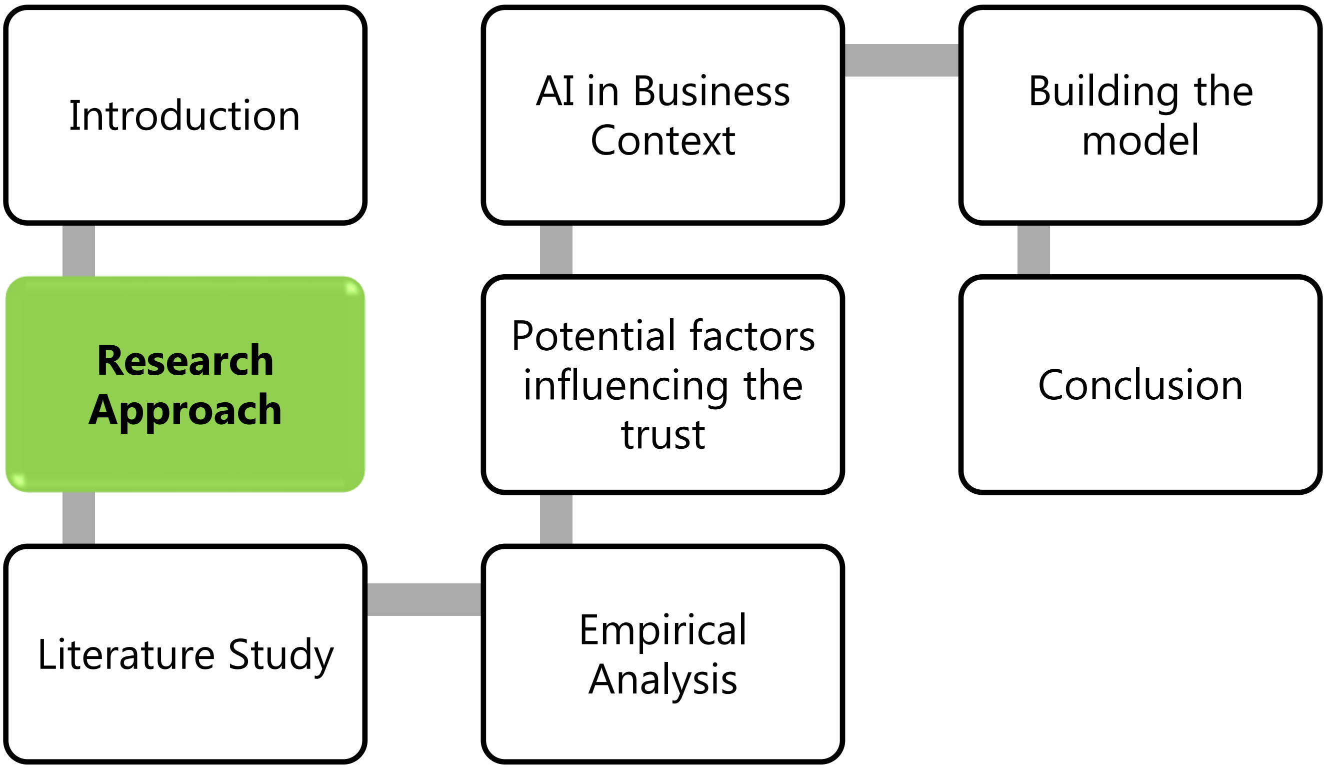

Are you consistent and complete with your results?
Are you being governed and are you auditable?
What data are you using, how are you using them , and is our data safe & secured?
Can you explain the decision that you make?
Will you be biased to any particular subjects / groups?
Are you trustable?
Can your results be interpretable?
DEVELOPMENT
Hello! I am an AI-based system. I have been assembled by a group of developers and have been programmed with the ability to learn, perceive, predict, and recommend. I am here to make your life easier.
Clients / potential users
Very interesting! But before adopting you, help us clarify our concerns and these concerns are to your developers/management.
Management (Directors, Managers)

Figure 5 - Actual challenges in adopting AI

2

RESEARCH APPROACH



Introduction

Research Approach

Literature Study

Empirical Analysis

Potential factors influencing the trust

AI in Business Context

Building the model

Conclusion

2.1 RESEARCH METHODOLOGY

The research is exploratory in nature in terms of determining the actual trust factors (data & AI model) and uses a qualitative study with an inductive approach by generating valuable theory. The research is mainly englobed with the literature review & desktop research, interview with the actors involved in the development of AI, and finally the use of a case study. In this research, the use of a qualitative approach can help in understanding the underlying perception of trust towards AI from the perspective of actors. The research can be divided into two main phases where the first phase involves the identification of potential factors that influence the trust in data and in the model and they are primarily derived based on the literature review & desktop research. The second phase involves the identification of trust factors from the perspective of actors involved in the development of AI. The factors identified in the first phase of the research would be used as the main source for the interviews with the actors. Using the literature and the findings generated from the interviews, a comprehensive model would be developed. This would further call for the evaluation which would be evaluated using interviews again and a single case study. Based on the reflections and insights retrieved from the case study and interviews, the final model would be developed and the initial working hypothesis that was laid would be validated at the end of the research.

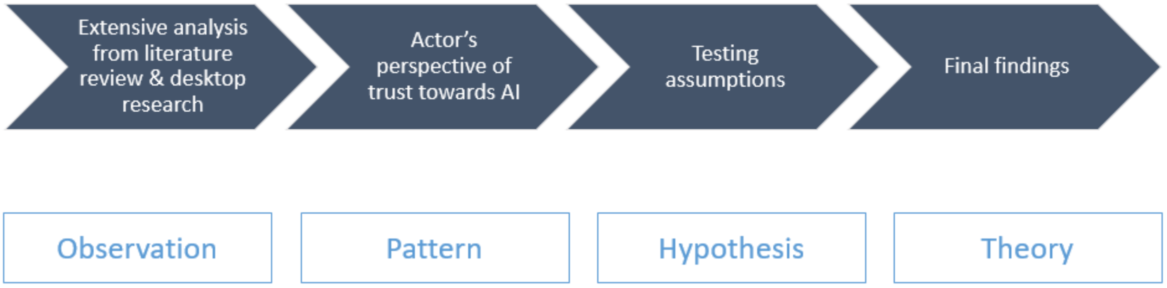

Extensive analysis from literature review & desktop research
Actor's perspective of trust towards AI
Testing assumptions
Final findings
Observation
Pattern
Hypothesis
Theory

Figure 6 - Building theory using the inductive approach

At this point, it would be appropriate to make a clear distinction on the actors involved in the development of AI, and these actors have to work together for implementing a successful AI application. The actors were identified based on PwC's current engagement with the clients for delivering AI-based solutions. However, not all actors mentioned below were involved but it would be apt to consider other relevant actors who would be involved in the coming years as the technology gets complex. The potential actors selected were based on communication with some of the AI experts and analysis done on several white papers. With this, the respondents for the interview would be the directors & partners, senior managers & managers, AI experts/specialist/researcher, data scientist, data engineer, risk advisors, and auditors. So, these would be the respondents who would be framed as "**actors**" in this context of the research.

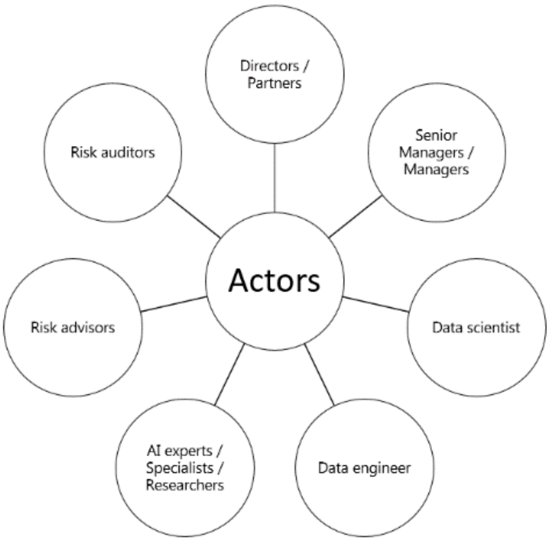


Figure 7 - Actors involved in the development of AI

2.2 RESEARCH FRAMEWORK

To understand the process involved in this research, a research framework would be developed that actually portrays a series of stages which would be followed to achieve the main objective of the research. Formulating a research framework can help in providing a clear picture of the research project. A set of an established pattern was recommended for formulating the research framework (Verschuren & Doorewaard, 2010). The phrases suggested by these researchers include

(1) Formulating the sources from which the perspective of the research would be developed

(2) Indicating to which research objects, the research perspective would be applied. (If there is more than one research perspective) , the third phrase has to be followed

(3) Indicating the ways, in which the analysis of the individual objects would be interrelated

(4) Stating the research project objective

This research would adopt the step by step approach proposed by the same researchers for constructing the research framework (Verschuren & Doorewaard, 2010). These steps include:

Characterizing the objective of the research project - The main objective of the research is,

To develop a model to assess and improve trust in Artificial Intelligence (AI).

Determining the objects of the research - The objects in this research are mainly the factors that influence the trust in AI. The factors in the context of the research are segregated as two sections: Trust factors towards the data and trust factors towards the AI model. The dimensions of DQ would be used as trust factors for the data.

Establishing the nature of the research - The research is qualitative and exploratory by nature and is mainly englobed with literature review & desktop research and interviews with the actors involved in the development of AI. More specifically, the research would develop a trusted AI model at the end which would contain the important trust factor specific to each phase in the model. These factors are identified and prioritized based on the studies from the literature sources, initial findings from the interview, and recommendations by the experts as part of the evaluation.

Determining the sources of the research – To have a clear perception of AI and trust, they would be studied separately initially in order to realize the value that can be provided to this thesis and then, they would be tied together by identifying the potential factors that would influence the trust in AI based on the concepts studied on trust with respect to automated systems, and robots retrieved from scientific literature studies. The potential factors identified from the scientific literature would lead the way for developing the conceptual model.

Making a simplified representation of the research framework – The simplified representation of the research frameworks are presented below:

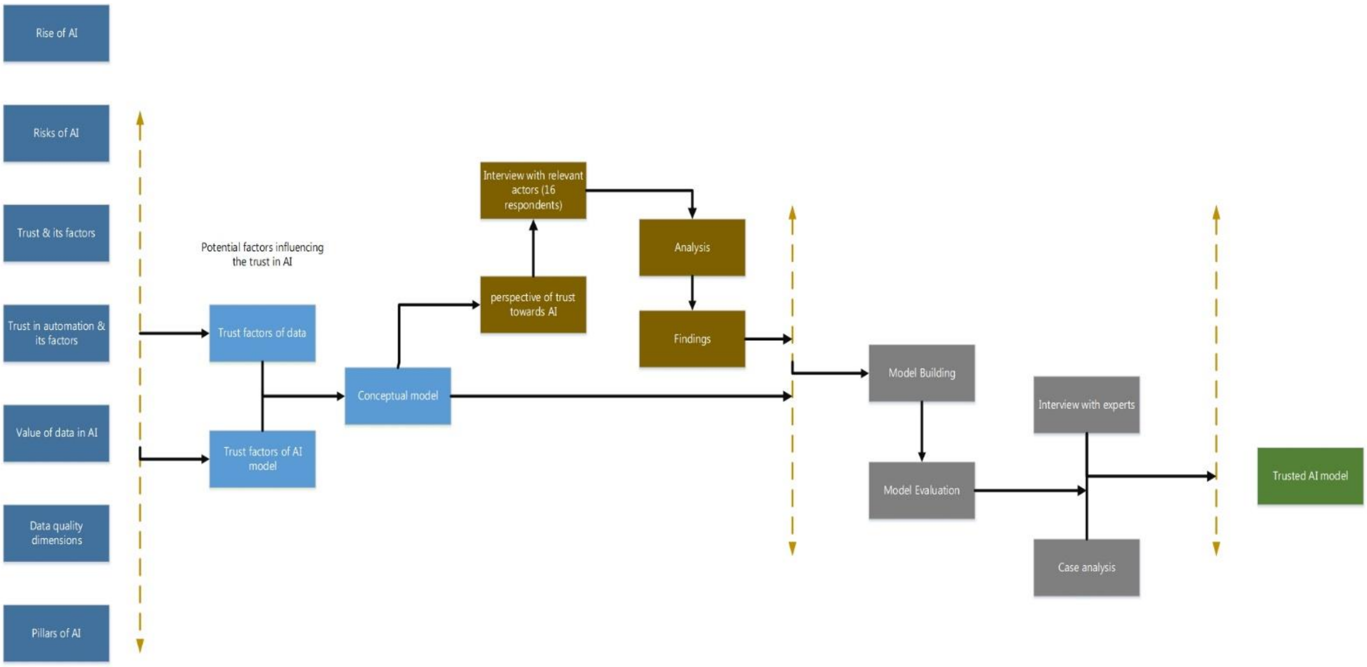


Figure 8 - Schematic representation of the research framework

Formulating the research framework elaborately according to the pattern defined above – An extensive analysis would be done on the concepts like trust & its factors in general, how trust plays an important role in automation, what are the factors that influence the trust in automation, how valuable is the data in AI, what dimensions would contribute to DQ, and finally what are themes that contribute to the pillar or core trust factors of AI. Based on the analysis of the literature, potential factors influencing the trust in data & in AI model would be identified. Based on the factors identified, a conceptual model would then be developed. The identified factors would serve as the main source for the interview. The main purpose of the interview is to understand the important factor from the perspective of actors involved in the development of AI and the questions for the interview were based on the factors that were identified from the literature source. Based on the analysis and findings from the interview, important factors towards the data and model would be identified and this will guide in building the initial version of the model. Since the model developed would be new and comprehensive, it would be further evaluated by conducting interviews but with fewer respondents (experts who are aware of the phases involved in the model) and with a case study. Combined with the initial analysis from the literature and reflections over the initial version of the model as a part of the evaluation, a final version of the trusted AI model would be developed.

2.3 RESEARCH QUESTIONS

In order to reach the prime objective of the research, the main research question and sub-research questions have to be answered. The questions that are framed will be answered as the research progresses. So, the main research question for this research is,

What are the factors of trust that influences the trust in Artificial Intelligence from the perspective of actors involved in the development of AI?

The below sub-research questions have been framed and answering these questions would actually provide clear inputs in answering the main research questions.

SUB-RESEARCH QUESTIONS

1. ***What are the various risks involved in AI? (Chapter 3)***

The issue of trust actually arises when a system either fails to accomplish the task or when the system possesses a threat to the environment. In the surroundings of AI, there are various risks and challenges that are actually concerning the users, business leaders in adopting and trusting AI. "With great potential comes great risk" would perfectly be tied in the situation of AI. The literature review & desktop research would be the prime input for determining the possible risks associated with AI.

2. ***What factors of trust influence the human-machine relationship in the current literature? (Chapter 3)***

When analyzing the concepts of trust in general, it would be flooded with a various number of factors seen from multiple dimensions of studies. Interestingly, not much of the research has been done on determining the factors that could possibly influence the trust in emerging technologies and AI, in particular. So, trust towards automation and their relevant models of trust were studied in detail. It can be assumed that these factors would hold the key for influencing the trust in AI as automation is considered to be one of the dimensions of AI. One of the primary sources of answering this question is through the use of literature. At first, trust in broader context along with the associated factors would be studied. Secondly, the emphasis would be placed on trust in the context of automation and what kind of factors would the human perceive in trusting such automated systems would be identified.

3. ***What are the dimensions of DQ that helps in improving the trust in the data? (Chapter 3 and Chapter 6)***

Although there has been a comprehensive study done on the dimensions contributing towards the quality of the data, not many researchers have highlighted its value in the medium of AI and what dimensions exactly contribute to DQ when AI is considered in the background till date. It is important to realize such dimensions, as data is believed to be the asset to the AI model & its resulting outcome. So, the quality of data is really a high priority that either a data scientist or data engineer would consider upfront. Combined with the analysis from the literature review and the insights derived from the interview would certainly help to answer this question. One potential blockage or a challenge is that there are multiple dimensions that can contribute to the DQ and the key is to figure out the most appropriate ones that can be considered with respect to the field of AI and big data.

4. ***What are the core themes (trust factors) of AI? (Chapter 3 & Chapter 7)***

As the importance of trust in AI has been in the top of the mind not just for the scientific researchers but also for many technology-based developers and providers and in fact, it has really been a stiff challenge for these technology-based companies to place trust in the AI systems that they develop, so that they can provide the same to the users. Many organizations and AI-based research institutions that have been established in recent years have started to confront the issues bounded around AI, in order to improve the trust. These institutions and organizations have established their own themes and principles for trusting AI. These themes would be identified at the first instance and later, it will be analyzed by ranking those themes in order to determine the core themes of AI. The literature review would be used as the main source for answering this question. However, It would be further analyzed at the end of the research to determine the actual factors of AI based on the findings generated from the literature review. It has to be noted here that, themes proposed by the institutions and companies would actually be considered as trust factors in this research.

5. *How does the model look like that influences the improvement of trust in AI? (Chapter 7)*

Based on the potential factors that were identified from the literature sources and based on the findings generated from the interview where the identified potential factors were used as an input, the model would be developed. At the first sight, the model would be expected to be comprehensive with several factors included and this would really call for the model evaluation by interviewing the right set of people and evaluating the model using a case study. Based on the reflections and feedbacks, the final trusted AI model would be developed. The model at upfront would establish certain prerequisites in order to realize the full potential of the trusted AI model. A complete demonstration of the model portraying them visually with detailed phases and subsequent factors with the indicators that are identified on each phase along with the prerequisites would certainly help to provide the answer to this question.

6. *What value does the trusted AI model provide? (Chapter 7)*

It would be essential to determine the value that the resulting model can provide in terms of scientific and real-world context post evaluation of the initial version of the trusted AI model and development of the final version of the trusted AI model. During the model evaluation, it would be expected to realize the actual usefulness of the model by interviewing with some of the relevant actors (Data scientist, Managers). In addition, few presumptions would also be made to further explore the possibilities in determining the value of the model in the real-world context.

2.4 GUIDING PRINCIPLES FOR QUALITATIVE RESEARCH IN IS

The research is mainly qualitative, it would tend to follow explicitly on some of the guiding principles upfront proposed by (Sarker, Xiao, & Beaulieu, 2013). These principles would aid the researcher to craft the research process systematically and ensuring that research is in line with the principles laid in order to provide successful qualitative research in the IS discipline. The principles that would be considered in this research would be the (1) The principle of internal coherence, (2) The principle of relevance, (3) The principle of theoretical engagement, (4) The principle of Transparency, (5) The principle of self-criticality.

The principle of internal coherence – Ensuring that different components presented in this research have a logical coherence and consistency.

The principle of relevance – It can be applied to various aspects like *relevance to the discipline, the relevance of the methodology, relevance to the practical world context*. At first, relevance to discipline – it is by ensuring that there is enough focus on the unique contribution associated with the **technology**. It has been said that most of the qualitative studies tend to focus on the social and behavioral issues, and often with technology bring no more than the context. The IS research would tend to lose its comparative advantage as compared to social or behavioral science research when there is not enough focus on the unique contributions associated with the technology in the IS research (Sarker et al., 2013). In this context of the research, the primary focus of technology is AI. The second aspect is with respect to the relevance of the methodology, is by choosing an appropriate qualitative research approach depending on the nature of the problem, and finally the practical relevance, which is done by ensuring that the research is grounded in reality by identifying the important factors from the perspective of relevant actors (in this research) and making the findings realistic and relevant.

The principle of theoretical engagement - Adopting the theory upfront into the research and developing a theoretical contribution at the end based on the outcome of the research. Additionally, ensuring that theoretical abstractions are being offered that results from the analysis and interpretation of the data. For example, the dimensions of DQ contributes to the data in the medium of AI applications.

The principle of transparency - To improve the value of the qualitative research, auditability and accountability of the work have to be ensured. For instance, it can be ensured by providing the details on how the interviews were conducted, who were the participants, how the data was analyzed, what were the inferences made, and demonstrating a systematic approach in arriving at conclusions from the data. Also, all the methodological assumptions, scope, and the procedures have to be made as lucid as possible.

The principle of self-criticality - Being watchful on the potential biases in the data sources during the data collection, potential flaws in the data analytics techniques and analyzing the data in a recursive manner to ensure that the anomalies between the data and perception of the researcher are eliminated to the maximum extent possible.

In addition to these guiding principles, some of the suggestions proposed by (Sarker et al., 2013) are being considered in this research. The recommendations are here as follows,

1. Research Focus component - Stating a clear research objective and research questions clearly upfront.

2. Methodology Component -

- (Data Collection) - Use of sampling logic when the study is conducted in an appropriate context. Suggesting the number of interviews reported and providing an outline/ guide for the interview process, Use of recording the interviews and transcribing would increase the credibility and auditability of the study.

- (Data Analysis) - Providing an adequate amount of coding details in the appendix can enhance transparency and thereby making the research more valuable.

3. Contribution component - Making a theoretical and practical contribution as that can make the overall research, more valuable.

4. Presentation component - Use of quotations can be seen as an impart to the level of richness in the qualitative study. Use of quotes can help to support facts, transfer the reader to that specific context and more importantly, it can easily convey the information, contribution, facts effectively to the readers.

2.5 DATA COLLECTION METHODS & ANALYSIS

As a part of data collection, different methods have been used in this research and the primary methods used are Literature review & desktop research, Interviews, and Case study.

METHODS

For this research, the qualitative approach has been taken as the research is mainly built based on the literature review, desktop research, interview with actors, and finally the use of a case study.

Literature review and Desktop research

Research search applications like Google Scholar, Scopus, Science Direct, and Research Gate had been the main source for accessing the insightful papers on AI, concepts of trust in general, and in automation. It was initially challenging to find papers related to trust in artificial intelligence from the management aspects as there weren't many articles emphasizing the importance of trust in AI or the factors that would influence the trust in AI. Since the articles of this category were considerably scarce, the method of backtracking was used to find some interesting papers by scanning the references of the fewer papers that were available. Also, extensive analysis was done on concepts of trust in automation & its factor which were actually used the main reference to identify the factors in the context of AI. So, popular papers on automation with respect to trust & its factors were studied. The literature review was done in an orderly process to improve the understanding of the concepts and to link the relationships between these concepts. Furthermore, several business articles, reports, white papers from several AI-based research institutions & tech companies have been used as a reference for this research. More specifically, to understand the important themes that can contribute to the essential trust factors of AI, business articles, white papers from several technology giants, papers published from AI-based research institutions were used as the main source for figuring out the important themes.

DATA COLLECTION SOURCE – LITERATURE REVIEW & DESKTOP RESEARCH

- Scientific Journals
- Books
- White papers & reports (Organizations)
- Conference papers
- Internet - Company Websites, business articles
- Papers from AI based research institutions

Figure 9 - Data collection source

Interviews

A guided, and a purposeful conversation between the individual or groups is called an interview (Sekaran & Bougie, 2016). They are done either face to face or by telephone or online. Interviewing the right set of actors involved in the development of AI and analyzing the data from those interviews is one of the prime aspects to this research as the model that would be developed is mainly based on the findings interpreted from the interviews apart from the insights derived from the literature sources. The interview in this context is done face to face and having a direct interview with the respondents would guide the researcher in adapting the questions accordingly and ensuring whether the responses received are properly understood by the researcher, and if not, the questions can be rephrased or reiterated in order to receive the desired response. The key aspect is to understand the important factors from the perspective of actors as it is believed that different actors would have a different perspective and more specifically, what a data scientist would see as an important factor to trust AI would be different from the perspective of a manager. The interview process would moderately be structured approach as the intention is to ask similar questions to the respondents. The techniques recommended by (Sekaran & Bougie, 2016) were adopted for improving the clarity of the questions and the overall interview process and these techniques could be found in chapter 6.1.

The approach of sampling was used to determine the right participants for the interview where sampling refers to the process of selecting a sufficient amount of people from the population as it could help to generalize the findings derived from those selected people to the whole population (Sekaran & Bougie, 2016). In this context of the research, it is necessary to interview and obtain information from the actors who are or who could be potentially involved in the development of AI. So, the method of **purposive sampling** is used which is actually confined to the specific types of people who can provide the desired information. Two phases of the interview are conducted, the first phase of the interview would be executed during the initial phase of the research and once the initial version of the model is developed, the second phase of the interview would be conducted but with fewer respondents which are mainly done to understand the feedbacks and reflections over the model.

Case Study

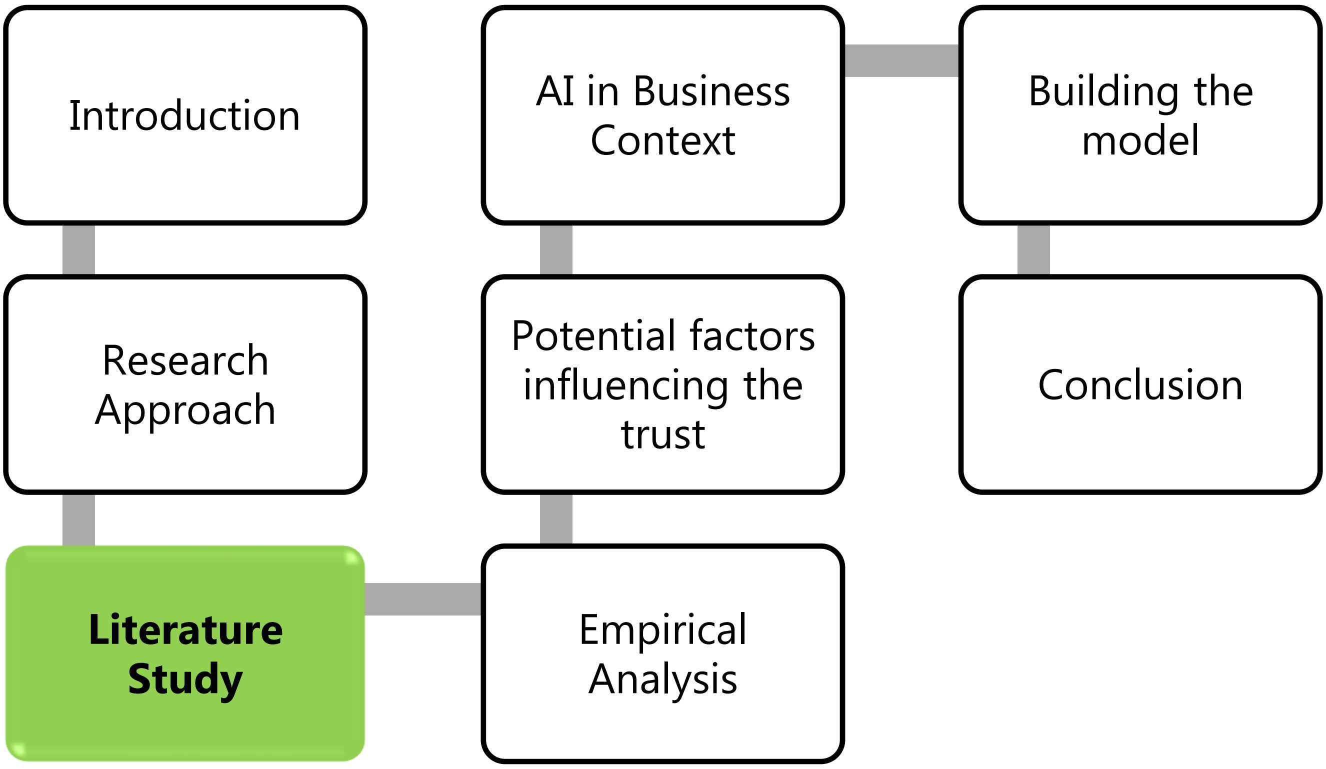
Another approach of data collection is through the use of a case study. Using case study can help in generating a theory which is a novel, testable and empirically valid theory. To generate such theories from the case study, a series of steps has to be followed as recommended by (Eisenhardt, 1989). The steps involved are: getting started, selecting the cases, crafting instruments and protocols, entering the field, analyzing within the case data, searching for cross-case patterns, shaping hypothesis, and enfolding literature. The main activity in the phase of getting started is to define the research questions and have some prior constructs upfront and these constructs are the trust factors defined in the initial phase of the research. This could provide a strong empirical grounding to generate a theory when the identified factors are proven to be crucial as the study progresses (Eisenhardt, 1989). One of the important aspects of the case study is selecting the right case for the research. In this study, a lone case was selected and the case was mainly used to evaluate the initial version of the model. The case involves the implementation of AI-based solution. Though the intention was to find more case with respect to the AI-based solution in the first instance and adopt a case study research methodology, it was rather hard in finding the suitable case with respect to this research or when any cases were found, not much of information was able to be gathered in order to proceed further. In the phase of crafting instruments and protocols, conducting an interview with one of the actors involved in the case was employed as the data collection methods and data were analyzed by manually transcribing the audio conversation to text. Realizing the fact that only one case was used to evaluate the model, it won't be appropriate to leap directly to the conclusion and one has to look for a cross-case pattern which can be through the existing literature or the interview done with other experts and actors falling outside the scope of the case. Based on the analysis and cross-case tactics, the hypothesis will be shaped. Finally, the comparison would be made with the existing literature which is the last and essential part of building theories from the case study.

ANALYSIS

It would be certain that one can expect a large amount of data while conducting the interviews and when it is being transcribed. A qualitative tool, called **NVivo** was used to analyze the data from the interviews. Although the tool had a feature of transcribing the audio recordings of the interviews to text. It was transcribed manually in order to respect the privacy of the individual and safeguard the confidential information if communicated any. The manually transcribed interview transcripts were then loaded into the qualitative tool and were coded in order to discard the irrelevant data and draw meaningful conclusions from the data. The tool was able to determine the commonality between every transcripts and theme using graphs. In addition to coding, a snapshot of every interview conducted was portrayed visually conveying the key findings as this could help the readers to understand the information easily from the interviews conducted and what kind of factors do these actors see as important in order to trust AI were presented in those snapshots.

3

LITERATURE STUDY



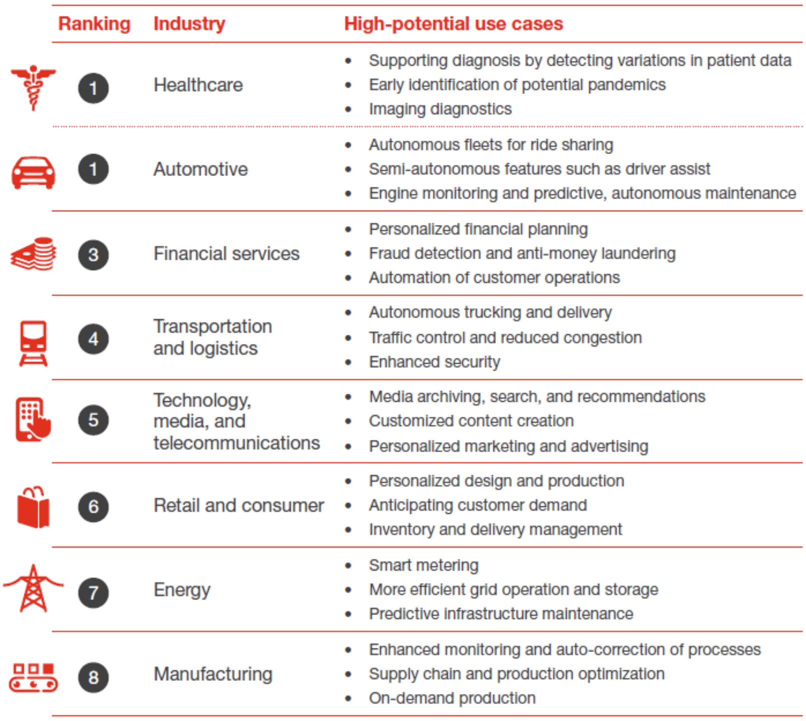
The main motivation of the literature study is to give a solid theoretical background on the concepts of Artificial Intelligence (AI), trust and their factors towards automation, human-robot interaction, the value of data & its dimensions and finally, the factors that are being considered as the core themes of AI. **At this point, it is important to note that the literature done on the concepts of trust in automation, human-robot interaction, etc. would serve as the main base in identifying the probable trust factors in the context of resulting AI model. Similarly, with respect to data, literature study done on the dimensions of DQ that were analyzed broadly and in the context of big data would be used as the main source of reference in identifying the probable dimensions of DQ in the medium of AI**. Although the concepts of trust were studied in general, neither its factors nor their models would be used in this research as the main purpose of studying them is to provide a solid background on the concept of trust & its factors and how it can lay as a starting point to study the concept of trust in automation. Finally, the literature study done on the themes of AI would be one of the valuable inclusion for this research in determining the essential trust factors mainly towards the AI model and these themes of AI are assumed as trust factors in the light of this research.

3.1 RISE OF ARTIFICIAL INTELLIGENCE

Entering into the era of Industry 4.0, business process and technological innovation would be driven by extreme automation and ubiquitous connectivity. The internet has completely changed the way we work, learn, socialize, creating innovative products & services, adopting new business models and reshaping the business markets and economy (Craglia et al., 2018). With the omnipresence of data and information sources, it has given rise to the evolution of digital technologies. Now, with the increased availability of computing processing power, exponential growth in the volume of data and algorithm has given foundation to the development of technologies like Artificial Intelligence (AI) (Craglia et al., 2018). To date, there is no agreed definition of what constitutes AI. It was in the 1950s in which Dartmouth research project introduced the concept of AI. According to (McCarthy, Minsky, Rochester, & Shannon, 1995), AI refers to

"*Ability of machines to understand, think, and learn in a similar way to human beings, indicating the possibility of using computers to simulate human intelligence* (McCarthy et al., 1995)".

Though the origin of AI has been there since the 1950s, it has captured immense attention over the last few years by realizing the value of the technology. In 2016, an AI-based computer program, "Alpha Go" developed by Google Deep mind had defeated eighteen-time world champion, Lee Seedol in an abstract strategy board game. This win by a machine was one of the triumphant moment for AI and it started gaining the attention of the public (Siau & Wang, 2018). Several applications of AI like Self Driving cars, Drones, Chatbots, etc. can provide enormous benefits if they are implemented correctly and these applications are still advancing & its encroachment is expected to intensify in the coming years. A number of countries have already put the self-driving vehicles, on the road and have started experimenting with driverless trucks for delivering goods and the same was being experimented with drones in delivering packages. One of the business advantages of AI is that they are highly scalable which can result in significant cost savings. Realizing the benefits, AI can cover a wide range of potential use cases across the industries. For instance, AI-based on-premise robots can bring in items to a consumer waiting in the dressing room. PwC has ranked several industries based on the possibility of deploying a practical AI to work and their potential use cases of having AI as an asset were identified in each industrial sectors.



Ranking	Industry	High-potential use cases
1	Healthcare	• Supporting diagnosis by detecting variations in patient data • Early identification of potential pandemics • Imaging diagnostics
1	Automotive	• Autonomous fleets for ride sharing • Semi-autonomous features such as driver assist • Engine monitoring and predictive, autonomous maintenance
3	Financial services	• Personalized financial planning • Fraud detection and anti-money laundering • Automation of customer operations
4	Transportation and logistics	• Autonomous trucking and delivery • Traffic control and reduced congestion • Enhanced security
5	Technology, media, and telecommunications	• Media archiving, search, and recommendations • Customized content creation • Personalized marketing and advertising
6	Retail and consumer	• Personalized design and production • Anticipating customer demand • Inventory and delivery management
7	Energy	• Smart metering • More efficient grid operation and storage • Predictive infrastructure maintenance
8	Manufacturing	• Enhanced monitoring and auto-correction of processes • Supply chain and production optimization • On-demand production

Figure 10 - Potential use cases of AI in industrial sectors, copied from (2018 AI Predictions: Responsible AI: PwC, 2018)

AI basically uses an external source of data and information from the Internet of things (IoT) and Big data as inputs for identifying the trends and pattern and makes a prediction, decision and recommendation accordingly (Kaplan & Haenlein, 2019). As it was indicated in chapter 1.1, AI can be classified into Weak AI and Strong AI. Weak AI, also called as Artificial Narrow Intelligence (ANI) focuses on doing the repetitive and narrow task. It is mainly specialized for a particular task. Strong AI or Artificial General Intelligence (AGI) can perform any cognitive function that a human can do and it can apply the intelligence to more than one specific problem (Ma & Siau, 2018). Having General AI is far from reaching its potential and the expectation is that it would take another couple of decades. Interestingly, IBM had defined one more category referred to as 'Broad AI' which is in between the Narrow and General AI. Broad AI is a collection of narrow AI systems that can make decisions and it's about integrating AI within a specific business process of an enterprise where one would require business & enterprise-specific knowledge, and data to train such systems (IBM, 2018).

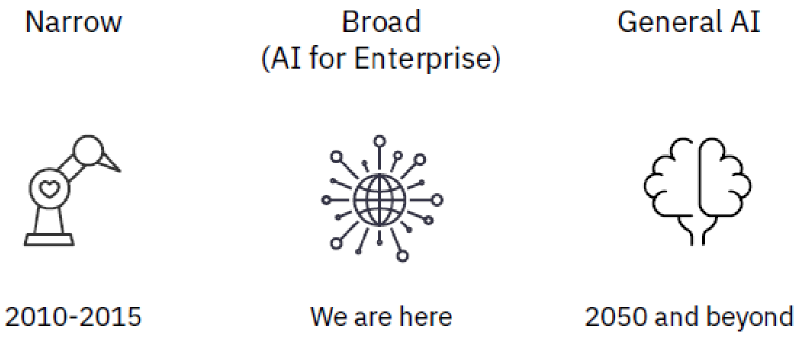


Figure 11 - Categories of AI, copied from (IBM, 2018)

As technology is advancing day by day, it has been said that AI has already surpassed human intelligence in some specific domains. These domains can be classified into 3 types based on the tasks of AI: General or Mundane tasks, Formal tasks, and Expert tasks. General or Mundane tasks could include visual & speech recognition, natural language processing, and translation, Formal tasks are related to games where some math's, theorem, and learning is involved, Expert tasks include tasks such as diagnosing disease, scientific & financial analysis and engineering (Garbhe, 2017). Although AI has gained its popularity in the recent years, there is still an ambiguity or confusion about the difference between AI, machine learning and deep learning but it can be simply stated as AI encompasses machine learning and deep learning. Figure 12 shows the clear distinction of AL, ML and Deep learning.

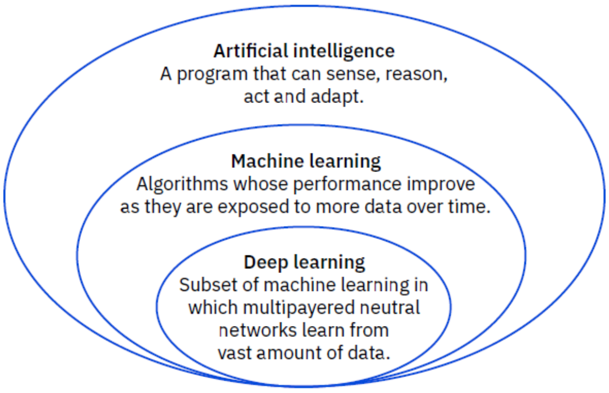

Artificial intelligence
A program that can sense, reason, act and adapt.
Machine learning
Algorithms whose performance improve as they are exposed to more data over time.
Deep learning
Subset of machine learning in which multipayered neutral networks learn from vast amount of data.

Figure 12 - Distinction between Artificial Intelligence, Machine Learning, Deep Learning, copied from (IBM, 2018)

There are several dimensions that AI can emerge, as the technology not only encompass ML and deep learning but also other dimensions like Robotics, Computer Vision, Speech and, expert systems, etc. that act as an umbrella for these technologies. The below figure gives a complete picture of various dimensions of AI & its subsequent applications.

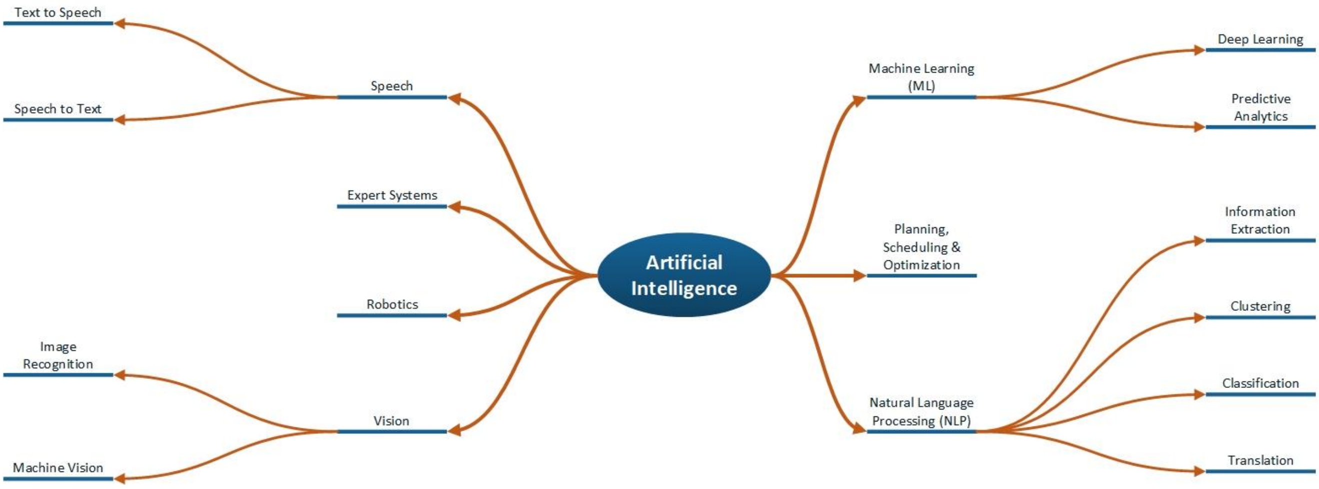


Figure 13 - Dimensions & Applications of AI

From all the dimensions mentioned above, Machine learning has been the buzz and it has been undergirding every aspect of the operations of big technology companies like Google, Facebook, Amazon, and Microsoft

(Bergstein, 2019). ML are machines that learn from the data without explicitly programming them with rules. Instead of programming the rules to the machine, it learns from the data that is being fed to the algorithms so that the algorithm can adjust accordingly to improve the accuracy of the algorithm (IBM, 2018). In essence, Machine Learning is the process of analyzing big data and applying the algorithm to fit the model accordingly in order to make a prediction (Sinnott, 2018). For example, in the retail sectors, ML are used to detect significant patterns in the present and predict the future, all it needs is the historical data on consumer behavior so that it can eventually predict how these consumers would behave going forward. A clear cut example of how ML works and the stages involved in ML has been elaborated in chapter 4.2 & 4.3.

3.2 RISKS OF AI

When Artificial Intelligence provides big opportunities for the business and society in terms of value and revenue, it can provide the same amount of risks when the technology goes wrong or misused. This section would identify & present an overview of risks encircled around AI. It should be noted that the section won't get into detailed or the technical aspects of all the risks that would be identified in this section.

To provide a clear context, the risks would be identified in terms of two categories: Data and model phase that one might encounter while dealing with AI as it is believed that there are several challenges or risks residing within these phases.

Data phase

Over recent times, there has been a growing concern about the importance of data. Several reports and studies have shown that the use of either non-representative or biased data could lead to unequal treatment of people based on gender, disability, ethnic origin, and religion (FRA, 2019). Many use cases where AI/ML models had shown biases in their results have made the society to question the reliability of the data with respect to the result. Instances of such use case can be seen as a large threat and such technology outcomes could land into massive jeopardy of risk to the society. (Crawford et al., 2019) presented a spotlight that there should be an urgent need of evaluation when commercial deployment of AI systems are used for prediction, classification, detection based on the race and the gender as it would cause a greater concern if those data in the systems are not properly evaluated. The quality of data & data as a whole in terms of representation is very important and if it has a poor representation or bad data quality, it can pose several challenges. For instance,

- Representation error - The data hasn’t covered all the population that is supposed to cover.
- Measurement error - The data hasn’t measured what they are intended to measure.

The data which is of low quality, outdated, incomplete, or incorrect at different stages of data processing would lead to poor prediction and assessment and in turn lead to bias, which would possibly result in breach of the fundamental rights of the individual, or purely incorrect conclusions and bad outcomes (FRA, 2019). Direct or indirect discrimination through the use of algorithms using data as the main source has been one of the pressing challenges. Examples of discrimination as a result of AI using inadequate data have been growing and has been a major worry. Instances like AI algorithm based recruiting tool preferring men over women (Details of this case is explained in chapter 1.2), Facial recognition system identifying more white people than the black people, machine translations showing gender bias, etc. The data that were used to train the AI / ML learning system are the reasons for the above outcomes. The data quality for building the algorithms and other AI-related technologies has become one of the major concerns for the fundamental rights compliant use of data. Following the frequently quoted principle “Garbage in Garbage out”, low quality of data would lead to low-quality outcomes from the model which can violate the fundamental rights. This doesn’t stop here, as low quality of data can affect the privacy and data protection which is another aspect of the risk (FRA, 2019). This has now fairly called into the concerns over the privacy and security of the data, since every individual, scientific experts are worried whether the data collected about them would be used unlawfully or unfairly against them (European Commission, 2019). As (Duursma, 2017) highlighted that data is the lubricating oil of AI systems and the user’s privacy is at stake at any event. People usually share information about themselves than they would if they are interacting with a person on the other side but when there are AI systems like chatbots on the receiver end of the information, it can remember every information that was communicated. In such an instance, people are worried about how data about them is collected and how their data has been used and whether their consent has been obtained (Microsoft, 2018). It's not just the privacy aspects that bother every individual of the targeted group but also the security of the data as every individual have their right to expect that their data is handled and stored securely. The security issue, for instance, can arise in the training data as the system should be able to distinguish between the malicious data and useful data before being fed to the system (Microsoft, 2018). (Davis, 2018) underlined the fact that data privacy & security has been a top priority since 2018 and in particular with EU's General Data Protection Regulation (GDPR) which has come into effect, the organization has to pay close attention with respect to compliance.

Every risk that is associated with the data is a chain of interconnected risks and it all emerges from the data and its quality that determines the positive and negative outcomes of the technology. Poor quality of data leads to a bad outcome and biased results which in turn calls for the concern over privacy & security of the data and finally questioning the transparency and governance of the data. When the data is not transparent, it would be hard to audit around the data process and this would certainly bring in the risk of auditability as well.

Furthermore, it is believed that today's economy is hinged upon the data as it can provide new values to the business and consumer but also can open doors to the dark web. Many organizations are using AI for cybersecurity purposes, but the hacker is also using the same technology to test their own malware in order to bypass the most advanced strategies (Woolley, 2019).

Model phase

One of the real risks of AI is that when the user or business don't trust the technology even when the technology provides enormous benefits. For instance, e-commerce website using an algorithm to provide a recommendation based on previous orders, the risk involved here would be low but what happens when an AI-powered algorithm turns down the mortgage applications without any reason or what if AI flags a certain individual at the airport security checks with no apparent justification. The users or the leaders may not trust AI in such instances if they can't understand on how it works and ideally, leaders won't be interested to invest in AI if they can't see evidence on how it made its decision (*2018 AI Predictions: Responsible AI: PwC*, 2018). The researchers (Wang & Siau, 2018) did also emphasized on the importance of AI being explainable as **low level of explainability & interpretability** can pose a significant risk to the users, developers, and governments even though AI can provide benefits for economic growth and social development. One of the researchers highlighted that the scariest part is not about the AI or the algorithm as they have been used for centuries but the automated decision making that AI can make without being able to explain the answer (Peters, 2018).

When AI is built using complex interconnected node systems (neural networks), they would be less capable of indicating the motivation or justification for decisions where only the input and output can be seen. Lack of transparency in this context is a major risk as every user either using or investing on AI would have underlying questions like what underlying thought has resulted in this kind of output?, What set of data were used to train the model?, How does the model think? etc. These questions have to be answered in order to trust the outcome of the decision made by AI.

(Boillet, 2018) classified some of the critical risks associated with AI especially when an individual or business adopts such technology and these risks involve Algorithmic bias, legal and liability risk, and reputational risks.

(1) Algorithmic bias: ML algorithm can identify the patterns or trends in the data and imagine if such patterns reflect some existing bias, the algorithm can amplify those biases which can produce outcomes that would reinforce the existing patterns of discrimination.

(2) Risk of cyber-attacks – AI systems could be easily targeted if the hacker wants to steal any personal and confidential information of an individual or an organization.

(3) Legal and Liability Risk – It is believed that there is little legislation governing AI and is set to grow in the coming years. AI systems using a large volume of consumer data has a stiff challenge in order to comply with the privacy regulation & EU's GDPR.

(4) Reputational risk – Realizing the fact that AI system can make critical decisions about individuals in a various range of areas by analyzing large volumes of data and if such system is biased, predicting bad results, being hacked, etc. it could pose a significant risk to organizations reputation that owns it.

Much to a broader context, PwC had identified six categories of AI risk with a varying impact not just on individuals but also towards the society and businesses. The six categories of AI include performance risks, security risks, control Risks, ethical risks, economic risks, societal risks.

NAME	DESCRIPTION	TYPES
Performance risk	As the algorithm ingest real-world data and preference as a source of inputs, they run the risk of learning and imitating human biases.	▪ Risk of errors ▪ Risk of bias ▪ Risk of opaqueness ▪ Risk of performance instability ▪ Lack of Feedback process
Security risk	Increasing concerns over the security of the model as well as the data when the systems get highly complex	▪ Cyber Intrusions risk ▪ Privacy risk ▪ Open-source software risk ▪ Adversarial attacks
Control risk	Organization adopting AI are expected to have clearly identified risks and controls	▪ Risk of AI going rogue ▪ Control malevolent AI
Economic risk	Possibility of impacting the jobs and shifting demands to different skills due to widespread adoption of AI	▪ Risk of job displacement ▪ Liability risk ▪ Risk of concentration of power within fewer companies
Societal risk	Possibility of developing "echo-chambers" between the humans and machine due to widespread adoption of AI	▪ Risk of autonomous weapon proliferation ▪ Risk of intelligence divide
Ethical risk	AI systems designed with a specific objective in mind which might compete with overarching organizational and societal values within which they operate	▪ Lack of value risk ▪ Value misalignment risk

Table 1: Categories of AI risk referred from (PwC's Responsible AI, 2019)

3.3 TRUST AND FACTORS OF TRUST

Trust could be termed as a notion that has gained immense attention especially, from various backgrounds of study – psychology, sociology, political science, economics, anthropology, history, etc. and in each study, trust has been approached with its own disciplinary ways and filters (Lewiki and Bunker, 1995). Most of the research done from various backgrounds have analyzed the position of trust in intermediating the relationship of the individuals, individuals & organization and, between the organizations. More importantly, trust has been examined as a crucial factor in interpersonal relationships focusing on romantic relationships (Rempel, Holmes, & Zanna, 1985). Trust has also been identified as a crucial factor in improving the productivity of the organizations and in strengthening the commitment of the organization (Nyhan, 2000). Similarly, in the research of exchange relationships, the main focus was on trust between the management and employees, supervisors and subordinates (Tan & Tan, 2000). (Butler, 1991), in his research, summarized the previous literature done on trust by stating that trust has emerged as an important aspect in interpersonal relationships and is very essential to the development of managerial careers, etc. However, attention towards the trust has grown immensely high over the last 10 years, as many have realized the importance of trust in promoting efficient cooperation and transaction and it has started to emerge as a prime focus of organizational theory, special issues of the Academy of Management Review, and International Journal of Human-Computer Studies (Lee & See, 2004).

Not surprisingly, many definitions of trust have been generated due to diverse interest from various disciplines. However, a degree of commonality exists among such definitions as the majority of trust definitions involve a trustor who would deliver trust, a trustee who would accept the trust, and a goal that needs to be achieved would be through a relevant behavior or efforts (Mayer et al., 1995). It was highlighted that some researchers see trust as an attitude or expectation and in that context, trust can be defined as in any of the following ways like: "*expectancy held by an individual that the word, promise or written communication of another can be relied upon*" (Rotter, 1967); "*expectation related to subjective probability an individual assigns to the occurrence of some set of future events*" (Rempel et al., 1985, p. 96); "*expectation of technically competent role performance*" (Barber, 1983); The above-mentioned definitions mainly deal with elements of expectation with regards to the behaviors or outcomes and it clearly indicates that trust concerns an expectation or attitude regarding the likelihood of favorable outcomes (Lee & See, 2004). Similarly, trust can be specified as an intention or willingness to act which means that trust is characterized as an intention to behave in a certain manner or to enter into the state of vulnerability. For instance, (Mayer et al., 1995) defined trust as

"Willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control the party (Mayer et al., 1995, p. 712)*."*

This definition by (Mayer et al., 1995) was considered as the most cited and accepted definition of trust in the academic studies and this definition of trust clearly projects vulnerability as a critical element of trust. Considering these definitions in mind, (Lee & See, 2004, p. 54) presented an elementary definition of trust which can be said as: "*Trust is the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability*".

Trust is always perceived as one of the important success factors for a business and it can be a major factor influencing various disciplines like capital investments, high tech development projects, relationship marketing, etc. (Blomqvist, 1997). Several experts and researchers had identified elements and factors of trust that could potentially influence the organization success. (Mayer et al., 1995) identified three factors as a basis for trust: Ability, benevolence, and integrity and these factors were regarded as a perceived factor of trustworthiness which was explained based on trust-trustee relationship.

⇒ Ability – a group of skills, competencies, and characteristics that enable the trustee to influence the domain.

⇒ Integrity – is the degree to which the trustee adheres to a set of principles the trustor finds acceptable.

⇒ Benevolence – the extent to which the intents and motivations of the trustee are aligned with those of the trustor.

Furthermore, a number of academic experts had discussed the similar factors influencing the trust, using relevant synonyms. For example, (Cook & Wall, 1980) also considered ability as an essential element of trust, while (Butler, 1991) used the word competence as a dimension of trust, which is still the same as ability. In the state of interpersonal relationships, trust is referred to as an attitude that is dynamic and follows a specific chain of dimension that forms moderately over time. Dimensions like predictability – the extent to which the future behavior can be anticipated (similar to ability) would form the basis of trust early in the relationship and it would be followed by dependability – the extent to which the behavior becomes consistent (similar to integrity) and as the relationship gets matured, the trust would lead to faith – the extent to which the person can be relied upon (similar to benevolence) (Rempel et al., 1985). So, the factors like dependability, predictability, and faith were considered as the major dimensions that influence the individual's acceptance of a trustee that forms the main base of trust. In the study of managerial trust, ten dimensions of trust were identified: availability, competence, consistency, discreetness, fairness, integrity, loyalty, openness, promise fulfillment, and receptivity (Butler, 1991).

The research won't get into the spotlight of trust factors derived from interpersonal relationships or relationships involving the humans in the process. Nonetheless, it would give the solid groundwork on some of the factors that are identified in this kind of relationships as it is presumed that some of these factors would be relevant in the context of automation or AI which would be seen in the following chapter 3.4. As (Adams, Bruyn, & Houde, 2003) underlined the fact that the need to trust automation arises from the same antecedents and factors that influences the trust in general. The below table presents an overview of the factors influencing the trust proposed by several academic researchers.

(Butler, 1991)	Availability, competence, consistency, discreetness, fairness, integrity, loyalty, openness, promise, fulfillment, receptivity
(Cook & Wall, 1980)	Trustworthy intentions, ability
(Mayer et al., 1995)	Ability, Benevolence, Integrity
(Rempel et al., 1985)	Predictability, Dependability, Faith
(Gabarro, 1978)	Openness, previous outcomes, Integrity, Discreetness, Motive, Judgement
(Hart, Capps, Cangemi, & Caillouet, 1986)	Openness/Congruity, shared values, autonomy/feedback
(Moorman, Deshpandé, & Zaltman, 1993)	Integrity, Confidentiality, Expertise, Tactfulness, Sincerity, Timeliness, Congeniality
(Giffin, 1967)	Expertise, Reliability as an information source, intentions, dynamism, personal attraction, reputation

Table 2: Summary of the trust factors retrieved from the literature study

3.4 TRUST IN AUTOMATION

Although not many models of trust were proposed in scientific academia with respect to building trust in AI models, several researchers had proposed models & frameworks in improving trust in automation. Automation in one way could still be considered as one of the initial categories of AI and since the term ‘’automation’’ may sound generic, some of the factors of trust could still play a major role in influencing the trust in AI.

One of the most widely accepted definitions of automation in the academic literature that summarizes the entire process of what automation is basically about was proposed by (Lee & See, 2004, p. 50),

“*Automation is a technology that actively selects data, transforms information, makes a decision, or control process*”.

The same researcher proposed a basic trust definition which is in line with their evaluation of trust as an attitude with respect to automation: “*the attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability*” (Lee & See, 2004, p. 54) and in this case, the agent could be an automated machine or an AI-based systems.

When automation becomes more complicated, the human operator’s ability to perceive the system becomes more diminutive. Though automation has great ability to improve the safety of the process and the resulting outcomes, the human operator is still dependent upon the functioning of the system, thereby creating a situation of vulnerability and uncertainty (French, Duenser, & Heathcote, 2018). Researchers have long argued that trust is a vital factor in intervening human-automation relationships. Normally, in the interpersonal relationships & other relationships involving humans, both the trustor and trustee would be the human but in the case of automation say for instance AI, the trustee would be the machine or the technology. The automation is likely to be accepted when trust in automation surpasses the human’s self-assessed ability to conduct a task but when the self-confidence surpasses the trust, the human worker is unlikely to use the automation (Lee & Moray, 1994).

Some of the academic foundations of trust formation were also realized in the automation literature and as (Rempel et al., 1985) proposed three dimensions: Predictability, dependability, faith as a basis for trust formation in their research on interpersonal relationship, these dimensions could be relevant with respect to automation as well. Predictability can form the initial basis of trust and its assessment mainly depends upon the actual predictability of the systems behavior, ability of the humans to assess the system predictability accurately, and the environmental stability in which the trust is occurring (French et al., 2018). If the system is transparent (ability to observe and understand easily) and could perform consistently as expected, it can produce predictability with a positive assessment to the humans over the system and can lead to the initial trust (Muir, 1994). In addition to predictability, (Lee & Moray, 1994) included automation reliability and ability as dimensions influencing at this stage of trust development. Once the initial trust is obtained, dependability would form the basis of trust after a period of time i.e., the degree to which human can rely upon the system. So, the quality of dependability is formed based on the gain of positively predictable behavioral experience, with the focus on event that involves risk or vulnerability and if it needs to be attributed towards automation, considerable amount of experience that system should have gained with a positive outcome in uncertain scenarios in order to be dependable towards the system (Muir, 1994). The human at this point would tend to trust the automation only if its process is complete and is able to achieve the human objectives within the pre-defined or operating environment (Lee & See, 2004). Faith would be the final phase in the trust formation model which is actually based on the future behavior of the trustee. Past predictability and dependability would be used as the basis for the belief that the trustee will behave in the future based on its past experience but in the case of automation, several processes involved in the automation are too intricate for humans to have a complete understanding of them and would require unexpected interaction (Muir, 1994). Similar progress emerged in the study of human operators adoption to a new technology where trust was dependent based on trial -and- error experience, followed by the understanding of the technology, and finally, faith (Zuboff, 1988).

Many factors influence the trust & their development towards automation and it changes over time and as a result, a number of theoretical models had been developed over the last twenty years. (Muir, 1994) initially developed a model for studying the trust in automation and it was based on the work done on interpersonal trust by (Rempel et al., 1985) as the model closely ties with the theoretical basis of trust foundation as mentioned above (Predictability, dependability, faith). Additionally, it was also presented that trust formation is mainly based on human expectations and they can be classified as persistence, technical competence i.e., the performance of automation over time, and fiduciary responsibility. In early 2000, (Kelly, Boardman, Goillau, & Jeannot, 2003) developed a trust model in automation which outlines the trust factors and their relationship between those factors. Three main references of trust were recommended: Understanding, the competence of the automation, and self-confidence. Understanding is laid on the explication of intention, predictability, and familiarity. Competence of automation is examined based on reliability, dependability, usefulness, and robustness. Self-confidence is developed based on skills, reputation, faith and practical experiences. (Hancock et al., 2011) developed a model where the factors of trust development in human-robot interaction were divided in terms of human-related, robot-related, and environmental categories. For instance, factors like dependability, reliability of the robot, level of automation, transparency, etc. were identified as performance-based factors in the robot-related environment while factors like anthropomorphism, the personality of robot, adaptability, etc. were identified as attribute-based factors in the same environment.

Perhaps, one of the most detailed and recent models that integrate the observational evidence on the factors that influence the trust in automated systems was produced by (Hoff & Bashir, 2015). The factors influencing the trust were organized into three categories: human operator, environment, and automated systems and the factors were divided into dispositional trust, situational trust, and learned trust. The dispositional trust would include the differences of the individual and human operator tendency which together forms the overall tendency to trust automation. Dimensions like Culture, age, gender, and personality traits contribute to dispositional trust (Hoff & Bashir, 2015). Situational trust includes a collection of internal and external factors relevant to the specific conditions under which the trust occurs and these factors will directly influence the trust in automation and determine the degree to which the trust influences the behavior towards automation. Finally, the learned trust involves the human operator's evaluation of the system based on past experience and present interaction with the automation (Hoff & Bashir, 2015). Furthermore, a trust model was designed in order to enable the inclusion of insights from behavioral trust theory in the design of automated systems (Hoffmann & Söllner, 2014). The model basically considers the antecedents of three dimensions underlying the formation of trust: performance, process & purpose (Lee & See, 2004). Performance dimension includes antecedents like competence, information accuracy, responsibility, and reliability over time while the process dimension contains factors like dependability, understandability, control, and predictability and finally, purpose dimension includes dimensions: motives, benevolence, and faith.

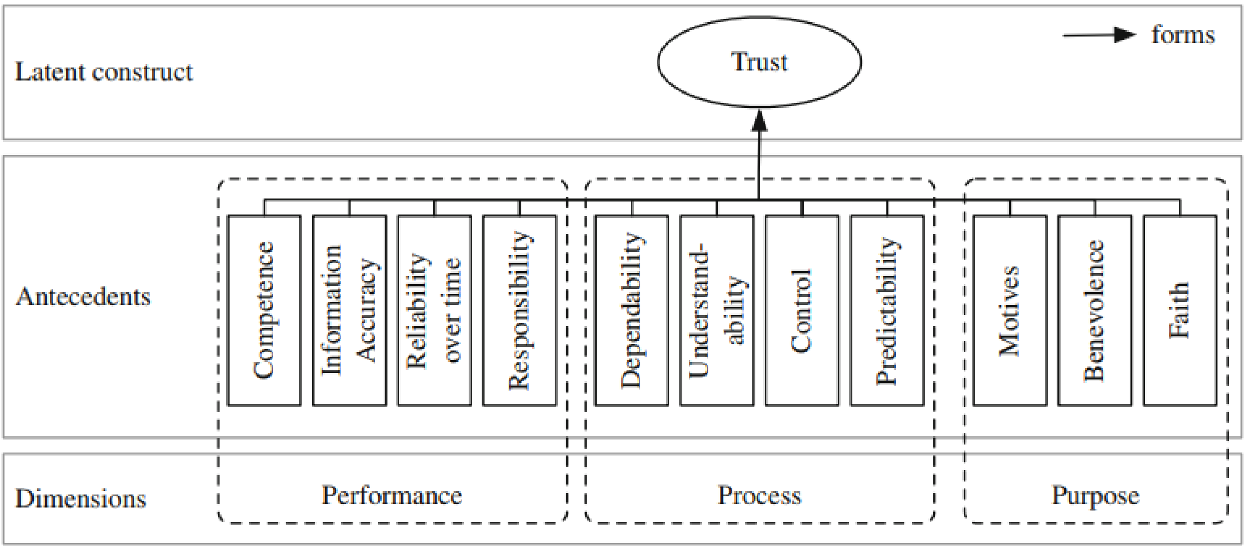
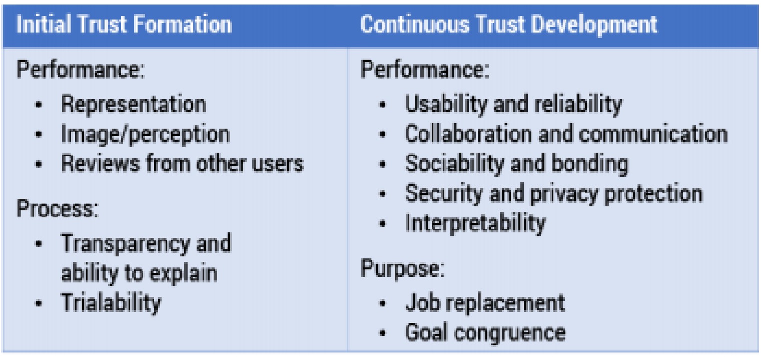


Figure 14 - Model of trust formation in automated systems, copied from (Hoffmann & Söllner, 2014), page 119

Very recently, (Siau & Wang, 2018) suggested some of the important factors of trust that are very crucial in building the initial trust and developing continuous trust in Artificial Intelligence (AI). Factors like representation, image/perception, reviews from other users, transparency & explainability, and trialability play an important role in building the initial trust towards AI. Once the trust is established, it needs to nurtured and continuously maintained where factors like usability & reliability, collaboration & communication, scalability & bonding, security & privacy protection, and interpretability would be the crucial factors to be considered that could help to develop continuous trust in AI (Siau & Wang, 2018).



Initial Trust Formation	Continuous Trust Development
Performance: • Representation • Image/perception • Reviews from other users Process: • Transparency and ability to explain • Trialability	Performance: • Usability and reliability • Collaboration and communication • Sociability and bonding • Security and privacy protection • Interpretability Purpose: • Job replacement • Goal congruence

Figure 15 - Features of AI that affects the trust-building, copied from (Siau & Wang, 2018), page 51

Normally, the factors that are influencing the trust in automation can be divided into three groups: human operator characteristics, automation characteristics, and environmental influences. There has been a great number of evidence where studies indicate that the characteristics of automation exert the largest effect on the level of trust. This was evident in one of the meta-analyses done by (Hancock et al., 2011) looking particularly at the automated systems like robots, found that the attributes of the robot and its performance were the largest characteristics to the development of trust followed by environment and human factors having a moderate & little effect towards the development of trust in automation.

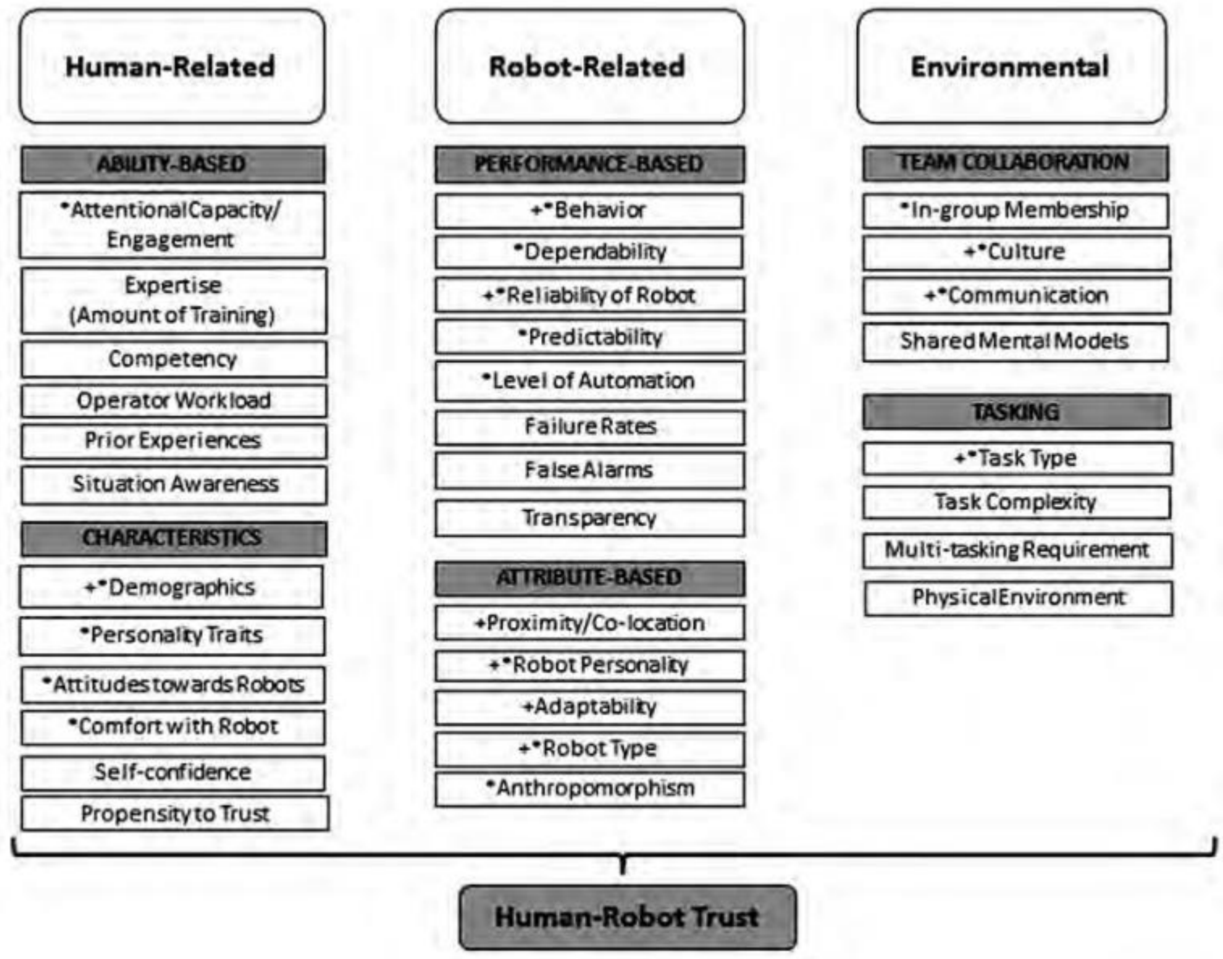


Figure 16 - Factors of trust development in human-robot interaction, copied from (Hancock et al., 2011), page 521

3.5 VALUE OF DATA IN AI

In an era of emerging technologies like Big data & Artificial Intelligence, every organization has to deal with a huge amount of data and bring out the value from them in order to make the technology trustworthy and reliable to the society (Côrte-Real, Oliveira, & Ruivo, 2017). Data is considered to be a critical input in any management practices across the industries. The exponential increase in the volume of data from people and businesses would have an enormous impact on how the organization approaches its critical process using such data's. AI typically includes a non-programmatic set of algorithms that thrive mainly on data. Advanced Artificial Intelligence and Machine learning mainly rely on data than the algorithm and if the data goes wrong, it can pose several risks like Data privacy & Security breach, Data leaks, and biases which might hamper the firm's reputation (Boillet, 2018). Many kinds of AI like machine learning, Deep learning would require an enormous amount of data which needs to be standardized, labeled, and cleansed of bias and anomalies. Otherwise, it would be ***garbage in garbage out*** process. The researchers (Ezry & Tyler, 2019) emphasized that a typical AI implementation requires a broad view of data that needs to be ingested as well as smart choices on data integration, governance, security, and privacy has to be made. It is believed that most of the enterprises are moving towards the data-driven decision making and it is absolutely essential to have a data which is not bad that could mean that all fields in the data should be present, there are no duplicates in the data, the data doesn't contain spelling or punctuation errors and should generally be correct (IBM, 2018). The outcomes derived from the bad data when fed to the AI model will lead to incorrect decision making. Researcher & decision-makers are gradually realizing the fact that a huge volume of information could bring benefits in understanding the needs of the customer, improving the quality of the service, predicting and preventing risks when one rapidly acquires and analyze the huge volume of data from various sources with various use. However, the use of data and its analysis are mainly reliant on ***the accuracy and the quality of the data*** which is absolutely necessary for generating the value from the data (Cai & Zhu, 2015). The quality of the data should be met with high standards regardless of the size and volume of the data as DQ is really the main foundation for the data to be trusted (Davis, 2018).

The notion of the quality of data not only depends on its intrinsic quality in terms of conformance to specification but also the actual use of the data in terms of conformance with the user's expectation (Janssen, Haryadi, Hulstijn, Wahyudi, & Van Der Voort, 2017). Furthermore, (Cai & Zhu, 2015) emphasized that the quality of data not only depend on its own features but also on the business process & its environment using the data and business users.

In the field of data quality, Researchers implicitly or explicitly had classified the data into three types: Structured data, unstructured data, and semi-structured data (Batini, Di Milano, & Maurino, 2009). The below table 3 shows its definition and examples:

Data Type	Definition	Examples
Structured data	Aggregation or generalization of items described by elementary attributes defined within the domain	Relational tables or Statistical data
Unstructured data	A generic sequence of symbols coded in natural language	Body of an email
Semi-structured data	Having a structure with some degree of flexibility	Markup language, XML

Table 3: Classification of data, referred explicitly from (Batini et al., 2009)

To improve the value of data, the approach of data-driven strategies has to be followed where a list of open-ended improvement techniques applied by data-driven strategies was recommended (Batini et al., 2009). Techniques include ***new data acquisition*** that improves the quality of data to replace the existing data quality problems, ***Standardization techniques*** that complements or replaces the non-standard data values with corresponding values that comply with the standard, ***Record lineage*** which identifies that data representation in two or multiple tables that might refer to the same real-world object***, source trustworthiness***, that selects the data sources on the basis of the data quality and , ***data & schema integration*** that would define the unified view of the data provided by heterogeneous data sources.

3.6 DATA QUALITY DIMENSIONS

Data is considered to be good only when it conforms to the relevant use and meet requirements and the best way to assess the quality of data is by using the Data quality dimensions. Data quality dimensions are commonly accepted and widely used methods in the assessment of data and it has been gaining increased attention from the researchers due to the emergence of AI and Big Data. Though one could find several dimensions of data quality in the literature, the research would consider only the relevant dimensions of the DQ as trust factors towards the data when it is being fed to the AI model.

Data and information are considered to be used interchangeably by most of the information system researchers. As in practice, managers and experts differentiate the information from data as information can be referred to as processed data. In this research, 'data quality' refers to both data and information. However, there would be clear differentiation made when stages like Data acquisition and Data preparation & validation terms are used in this research and in such case, it would be stated explicitly.

Data Quality (IQ) has been one of the prime concern for organizations and it has become an active area of Management Information Systems research (MIS) (Lee, Strong, Kahn, & Wang, 2002). The exponential growth of data and direct access to information from multiple sources by the users have increased the need and awareness for high-quality information in organizations. The meaning of DQ lies in how the data is perceived and used by the user and it is basically the perception for the user that defines DQ (Miller, 1996). The Total Data Quality

Management group of MIT University led by Professor Richard Y.Wang, who has done extensive research in the data quality areas, defined "data quality dimension" as a "**set of data quality attributes that represent a single aspect or construct of data quality"** (Wang & Strong, 1996, p. 6). The researcher developed a conceptual framework of data quality using an empirical approach that identifies four categories containing fifteen DQ dimensions. The main four categories were Intrinsic DQ that denotes that data has their qualities in their own right, Contextual DQ tells about the requirements that DQ must consider within the context, Representational DQ and accessibility DQ emphasize the role of the systems. Dimensions such as believability, accuracy, reputation, and objectivity were considered to be the most fundamental part of Intrinsic DQ while dimensions like relevancy, timeliness, completeness, appropriate amount of data, etc. were related to Contextual DQ. As representational DQ relates to the format and meaning of data, dimensions such as interpretability, ease of understanding, and concise representation were a part of representational DQ.

A theoretical approach was followed by (Wand & Wang, 1996) to define the DQ dimensions where the approach considers the information system (IS) as a representative of the real-world system (RW) (Wand & Wang, 1996). RW is said to be properly represented in an IS only if there exists a complete mapping RW -> IS and no two states In RW can be mapped into the same state in IS.

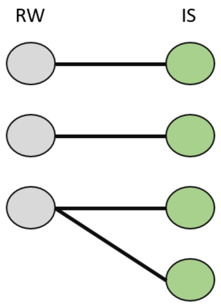


Figure 17 - Proper representation of a real-world system, referred explicitly from (Wand & Wang, 1996)

Based on the proper representation definition, (Wand & Wang, 1996) identified three design categories of deficiencies: incomplete representation, ambiguous representation, and meaningless state of representation and based on this deficiencies, a set of DQ dimensions were defined with reference to those deficiencies. More specifically, the identified DQ dimensions were: accuracy, reliability, timeliness, completeness, and consistency (Wand & Wang, 1996). The overview of these dimensions would be found in Table 4.

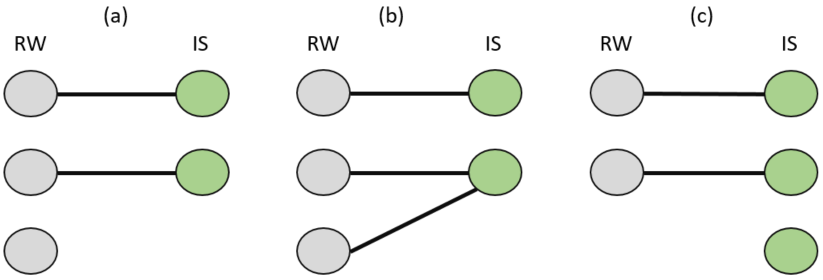


Figure 18 – Incomplete (A), Ambiguous (B), Meaningless (C) representations of a real-world system, referred explicitly from (Wand & Wang, 1996)

(Batini et al., 2009) highlighted that dimensions like accuracy, completeness, consistency, and timeliness were regarded as a basic set of DQ dimensions and has constituted the focus of the majority of researchers assessing the quality of data. Furthermore, a conceptual framework was developed for assessing the DQ by (Bovee, Srivastava, & Mak, 2003) where the model consisted of four main attributes: accessibility, interpretability, relevance, and integrity. To evaluate the integrity of the data, four sub-attributes were identified: accuracy, completeness, consistency, & existence. These attributes relating to the integrity are **intrinsic** and it relates to the process of how information was created while attributes like accessibility, interpretability, and relevancy are **extrinsic** in nature.

The ubiquitous amount of data and information led to the rise of big data in recent years, which has made the researchers propose a universal two-layer big data quality standard for assessment where five DQ dimensions its associated elements were identified (Cai & Zhu, 2015). DQ dimensions like availability, usability, reliability, and relevance were considered as important and inherent features of the data quality.

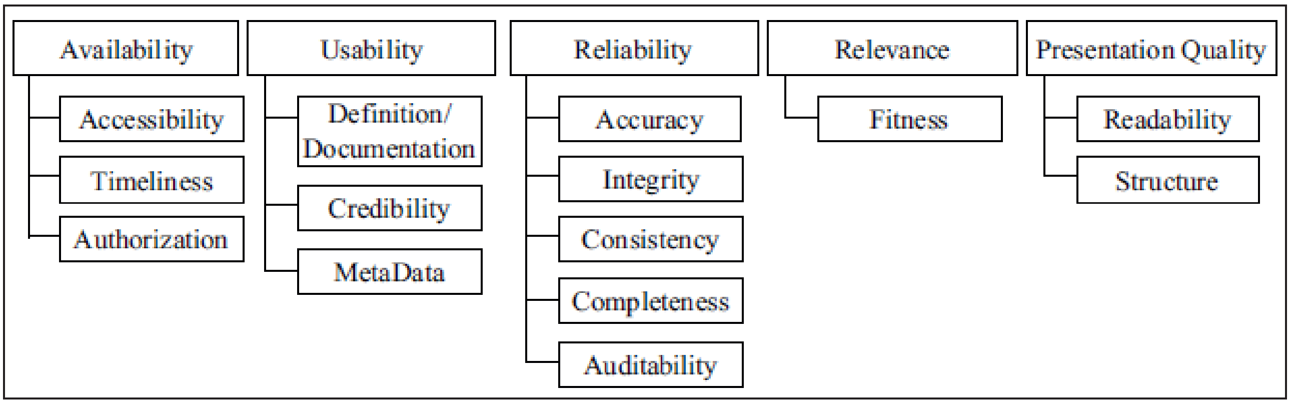


Figure 19 - A universal, two-layer big data quality standard assessment model, copied from (Cai & Zhu, 2015), page 4

Overview of the DQ dimensions

As (Janssen et al., 2017) stated that there has been no consensus on what constitutes the dimensions of DQ as there are several dimensions found in the academic literature. Also, there is no exact meaning of each dimension of DQ. However, it has been made to ensure that the meaning of each dimension retrieved from the literature source provides relevancy to this research and conveys the same interpretation when such dimensions are used in the context of AI.

DIMENSIONS	DEFINITION	SOURCE
Accuracy	The extent to which the data is correct, reliable and certified. Data is accurate when data values are stored in the database that corresponds to real-world values. The degree to which the information reflects the underlying reality. The extent to which the information is true or error-free with respect to some known, designated or measured value. Information that represents the real-world state.	(Wang & Strong, 1996), (Batini et al., 2009), (Miller, 1996), (Bovee et al., 2003), (Janssen et al., 2017)
Consistency	The extent to which the data is present in the same format and could be compatible with the previous data. The extent to which the multiple recordings of the value for an entity attribute is the same or closely similar across time or space. The information (processed data) and data source have no contradiction.	(Pipino, Lee, & Wang, 2002), (Bovee et al., 2003), (Janssen et al., 2017)

Completeness	The extent to which data are not missing and are of sufficient breadth and depth for the task at hand. Data having all the required parts of entity information. The ability of the information system to represent every meaningful state of the represented real-world system. Information having all the parts of the entity description. The degree to which all possible states are represented in the real-world state	(Pipino et al., 2002), (Batini et al., 2009), (Wand & Wang, 1996), (Bovee et al., 2003), (Janssen et al., 2017)
Accessibility	The extent to which the data is available, or easily and quickly retrievable. The extent to which the information can be obtained when accessed. Ability to get the information which sounds useful.	(Pipino et al., 2002), (Miller, 1996), (Bovee et al., 2003)
Timeliness	The extent to which the data is sufficient for the task at hand (Up to date). The delay between a change of real-world state and the resulting modification of the information system state. Availability of the data on time and how up to date the information is.	(Pipino et al., 2002), (Wand & Wang, 1996), (Janssen et al., 2017)
Security	The extent to which the data access is restricted in order to maintain its security. The degree to which the information is protected from people (logical security) and protecting information from natural disasters.	(Pipino et al., 2002), (Miller, 1996)
Believability	The extent to which the data is regarded as credible. The extent to which the information is regarded as true and credible.	(Pipino et al., 2002), (Wang & Strong, 1996), (Janssen et al., 2017)
Relevancy	The extent to which the data is regarded as applicable and helpful for the task at hand. The extent to which the information is addressing the actual needs. The degree to which the information is applicable to the domain and the purpose of interest in a given context. Information that addresses customer needs.	(Pipino et al., 2002), (Miller, 1996), (Bovee et al., 2003), (Janssen et al., 2017)
Interpretability	The extent to which the data is in appropriate language, symbols, unit and has clear definitions. Ability to understand the information and find the meaning in it.	(Pipino et al., 2002)(Wang & Strong, 1996), (Bovee et al., 2003)
Reliability	The extent to which the data is correct and reliable. Whether the data can be counted to convey the right information	(Wang & Strong, 1996), (Wand & Wang, 1996)
Availability	The extent to which the data is physically available	(Knight & Burn, 2005), (Wang & Strong, 1996)
Auditability	The extent to which data accuracy and integrity can be evaluated within the rational time and manpower limits during the data use phase	(Cai & Zhu, 2015)
Usefulness	The extent to which the information is applicable and helpful for the task at hand	(Wang & Strong, 1996)

Table 4: Overview of the dimensions of DQ

3.7 CORE THEMES (TRUST FACTORS) OF AI

The trust issues in AI is top of mind for several academic researchers, experts, technology developers, users and society as AI can exhibit vulnerabilities such as privacy & security concerns, lack of explainability & transparency, exposure to bias, etc. (Mojsilovic, 2018). Academic researchers & experts are confronting such issues with a scientific approach by developing techniques, methods, algorithm to assess and address the pressing issues around AI. Nevertheless, it is also critical to lay down some of the foundational elements of trust as themes of AI in order to trust the system at the end. These foundational elements could help the organization to assess the technology based on the elements defined.

There is no consensus or universally accepted elements on what contributes to the foundational elements of trusted AI or the themes of AI as several research institutions and technology giants like Facebook, Google, IBM, and, Microsoft have defined their own principles and themes of AI.

For instance, IBM Research has proposed several elements that form the basis for trusted AI systems and they are namely Fairness, Robustness, Explainability, and Lineage (Mojsilovic, 2018).

THEMES	DESCRIPTION
Fairness	Using training data and models that are free of bias in order to avoid unfair treatment of certain groups
Robustness	AI systems should be safe and secure and not vulnerable to tampering or comprising that data that is being trained on.
Explainability	Providing decision and suggestions that can be understood by their users or developers
Lineage	Including the details of the development, deployment, and maintenance in order to be audited throughout the life cycle.

Table 5 - Pillars of AI proposed by IBM, referred from (Mojsilovic, 2018)

While Microsoft did propose some of the AI principles that would help them to design trustworthy AI systems as the principles defined below should be rooted in important and timeless values.

THEMES	DESCRIPTION
Fairness	Making sure that AI systems are treating all people fairly.
Reliability & Safety	Making sure that AI systems are performing reliably and safely.
Privacy & Security	Making sure that AI systems are secured & are respecting privacy.
Inclusiveness	Making sure that AI systems are empowering everyone & engaging people.
Transparency	Making sure that AI systems are understandable.
Accountability	Making sure that AI systems have algorithmic Accountability.

Table 6 - Microsoft principles on AI, referred from ("Microsoft AI principles," 2019)

In order to establish Responsible AI practices, Google recommended practices highlighting the importance of building fairness, interpretability, privacy, and security into the AI systems ("Responsible AI Practices," 2019).

THEMES	DESCRIPTION
Fairness	➢ Designing a model having concrete goals for fairness and inclusion. ➢ Using representative datasets to train and test the model. ➢ Checking the system for unfair biases. ➢ Analyzing the performance of the system.
Interpretability	➢ Planning out the options to pursue interpretability. ➢ Treating interpretability as a core part of the user experience. ➢ Designing a model that is interpretable. ➢ Understanding the trained model.
Privacy	➢ Collecting and handling data responsibly.

	➢ Leverage on device process where appropriate. ➢ Safeguarding the privacy of AI/ML models.
Security	➢ Identifying the potential threats to the system ➢ Developing an approach in order to combat the threat.

Table 7 - Responsible AI practices by Google, referred from (“Responsible AI Practices,” 2019)

Ensuring that AI systems are trustworthy and responsible has become one of the top priority & Challenges not only for technology giants like Google, IBM, Microsoft but also for technology advisory based companies like PwC, Deloitte, EY, etc. For instance, PwC had identified five main aspects of responsible AI in order to build trust to the stakeholders (*Building Trust in AI and Data Analytics*, 2018). Interestingly, all the principles laid by big tech companies like Google, Microsoft, IBM, etc. were focused towards the outcome of the model and the model itself but companies like BigFour (PwC, Deloitte, KPMG, EY) are ensuring that AI systems developed have proper governance, regulatory standards, and ethics which is something a paramount for trusting AI.

THEMES	DESCRIPTION
Fairness	Whether the bias in the data and AI model is minimized and Are the bias been addressed when AI is being used?
Interpretability	Can we explain how an AI model is making decisions and ensuring that the decisions made by AI models are accurate?
Robustness & Security	Can we rely on the performance of the AI system and whether AI systems are vulnerable to attack?
Governance	Who is accountable for the AI system and ensuring whether proper controls are in place?
System Ethics	Whether the AI system is complying with the regulation and how will AI system impact the employees and customer?

Table 8 - Five dimensions of Responsible AI recommended by PwC, referred from (*Building Trust in AI and Data Analytics*, 2018)

Recently, A high-level expert group on AI set up by the European Commission published a document on “Ethical guidelines for trustworthy AI ” with the objective of promoting trustworthy AI (European Commission, 2019). The document has highlighted that Trustworthy AI contains three components: **Lawful** - AI complying with all the applicable laws & regulation, **Ethical** – ensuring adherence to ethical principles and values, **Robust** – AI being robust from a technical and social perspective, even with good intentions. Furthermore, seven key requirements for trustworthy AI were established in order to ensure that development, deployment, and use of AI are meeting those requirements.

THEMES	DESCRIPTION
Human agency and Oversight	Including fundamental rights, human agency, and oversight
Technical robustness and Safety	Including resilience to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility
Privacy and data governance	Including respect for privacy, quality, and integrity of data, and access to data.
Transparency	Including traceability, explainability, and communication
Diversity, non–discrimination, and fairness	Including the avoidance of unfair bias, accessibility, and universal design, and stakeholder participation.
Societal and environmental well being	Including sustainability and environmental friendliness, social impact, society and democracy
Accountability	Including auditability, minimization, and reporting of negative impact, trade-offs and redress

Table 9 - Requirements for building a trustworthy AI, referred from (European Commission, 2019)

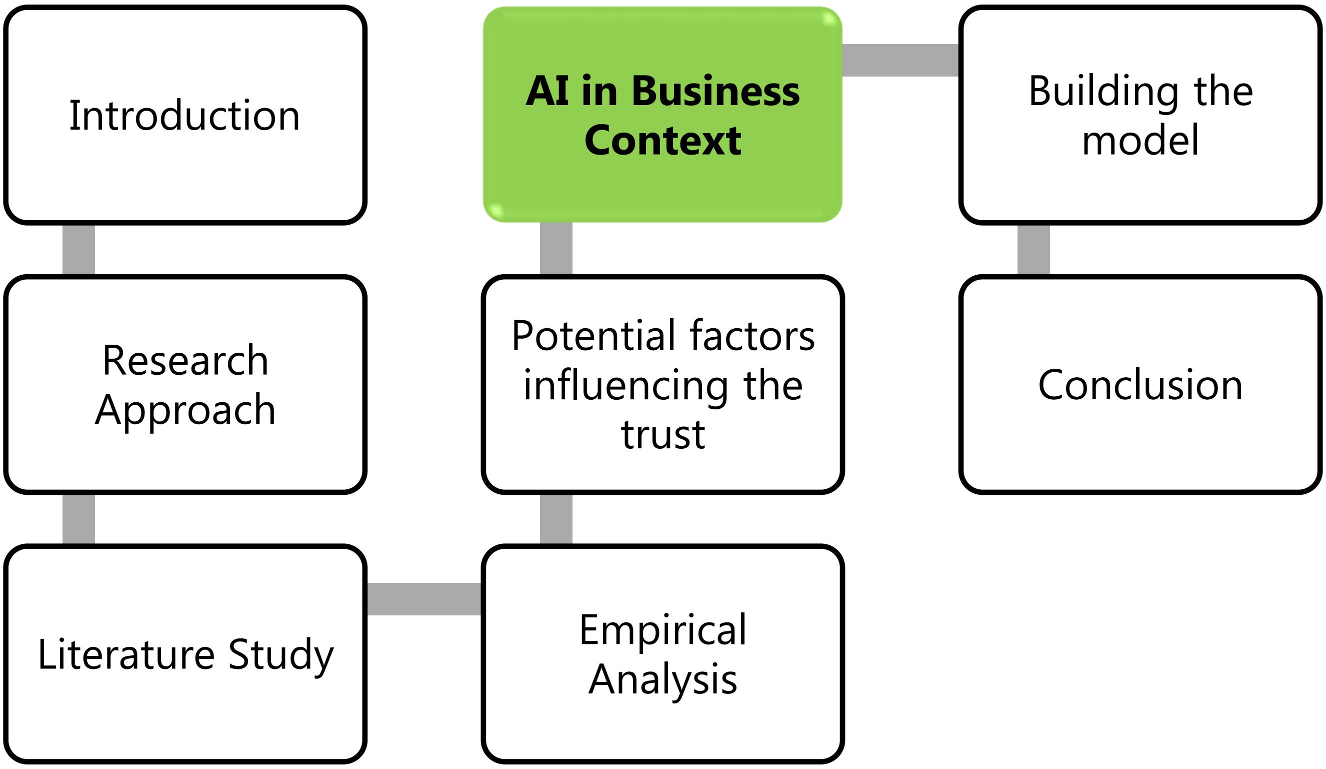
Additionally, several research institutions have now been emerged with the main motive to educate the public and ensure that technologies like AI, robots serve humanity in a beneficial and responsible way. These research institution's main theme is to address the key challenges like Biases, Safety, Inclusion, fairness, privacy-related issues, and transparency, etc. The below table gives an overview of research institutions and their main themes or the principles in order to build AI for the benefit of all.

NAMES OF THE RESEARCH INSTITUTIONS	THEMES
Partnership on AI, ("The Partnership on AI," 2018)	▪ Safety-Critical AI ▪ Fair, Transparent, and Accountable AI ▪ AI, Labor, and the Economy ▪ Collaboration between people and AI systems ▪ Social and Societal Influences of AI ▪ AI and Social Good
AI now Institute, ("AI Now Institute," 2017)	▪ Rights & Liberties ▪ Bias & Inclusion ▪ Labor & Automation ▪ Safety & Critical Infrastructure
The Institute for Ethical AI & Machine Learning, ("The Institute for Ethical AI & Machine Learning," 2018)	▪ Human Augmentation ▪ Bias Evaluation ▪ Explainability by justification ▪ Practical accuracy ▪ Trust by privacy ▪ Security ▪ Displacement Strategy ▪ Reproducible Operations

Table 10 - Core themes laid by several AI research institutions

4

AI IN BUSINESS CONTEXT



4.1 ACTORS INVOLVED IN AI DEVELOPMENT

In a typical AI journey, one would assume that the main stakeholders involved would be the technology investors which could be the clients, technology creators would be the companies developing the AI-based systems or solutions and the technology users which could be the consumers or the client itself. However, the context would vary as in some case the technology creators and investors would be the same. For example, Amazon creating AI-based products to the consumer and in this case, technology investors & creators are Amazon while the consumer is the users of the technology.

In the case of technology creators, it is essential to identify the main actors involved in the development of AI-based solutions as most of the literature and business articles emphasize only the actors like Data Scientist, Data Engineers, and developer as the core actors in developing the technology. More than that, it needs to be realized that there are several other actors who could really influence technology. The identification of actors here, in this case, has been made in partnership with PwC consultants in order to realize the current stakeholders and actors involved in their development of AI-based solutions.

It was understood that actors like Directors, Partners, and the senior managers from Data analytics department would be people who would start to engage with the client at the first instance, understand the client background & their business models, look for the improvement needs and this would be the scenario for most of the case. There are also possibilities where the clients directly approach PwC for a solution or they would request PwC to develop the AI model if they already know the solution for the problem but lacks in house technical capabilities. In either case, actors like Directors, Partners, Senior Managers, and Managers would be the ones who would have direct relationships with the client.

Directors & Partners - are the people who are mainly responsible for signing the initial proposal plan and they would be involved in the first kick-off meeting with the clients and the managers. They have the ultimate responsibility for the whole project right from the initial plan to the final deliverables. Once the model has been developed, they had to provide sign off on the final deliverables and give the final presentation to the clients about the deliverables.

Senior Managers & Managers - are mainly responsible for selling the engagements, making the proposals by working with the team, contacting the clients & providing regular updates to the client as a part of the deliverable, managing the engagement & the team by tracking hours and escalating things if necessary, scheduling the important meetings with the client.

Senior Associates & Associates (Data Scientist, Data Engineer, DevOps Engineer) - are the ones who would be performing the core work in the engagements. In the context of AI, this would include right from gathering the data to deploying the model. However, there is always a misinterpretation on the roles of a Data Scientist and Data Engineer. So, who is called as Data Scientist and who are data engineers and what roles & responsibilities make them differentiated? (Willems, 2017) gave a clear differentiation on both these actors,

A data engineer is someone who would:

- ➢ Discover opportunities in getting the data.
- ➢ Recommend ways to improve data reliability, efficiency, and the overall quality of the data.
- ➢ Developing data set processes for data modeling, mining, and production.
- ➢ Work with database systems and tools for ETL purposes.

A data scientist is someone who would:

- Conduct research to answer the business & industry questions and concerns
- Makes use of a large volume of data to find an answer to the business.
- Exploring the data to find patterns and trends.
- Builds predictive and prescriptive modeling based on the data.

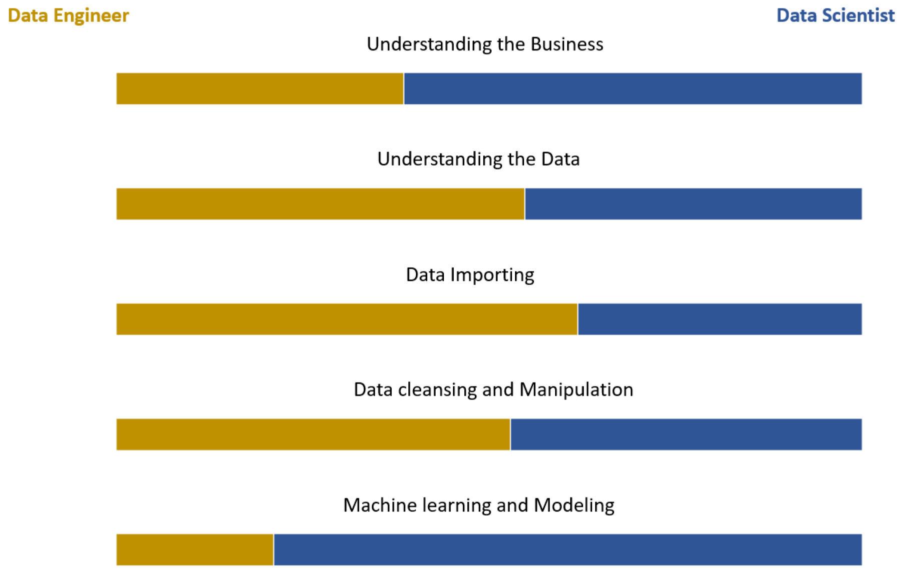


Figure 20 - Workflow of a Data Engineer & Data Scientist, referred explicitly from (Willems, 2017)

It would be effective only when both the actors work together to get the best value from the data and provide insights & answers to business-critical decisions (Willems, 2017). However, by having interaction with several employees at PwC especially from Data Analytics department, It was recognized that there is no specific role segregation currently as to who works on data, who builds models, and who deploys the model as most of the times it could be the case where a couple of data scientists would be responsible for all the tasks involved in the development of AI projects. In such cases, the stakeholders would be the partner & directors from the top management level, Managers from the middle management and the Data Analytics people (Data Scientist, Data engineer, etc.). The below-mentioned figure 21 gives an overview of the stakeholders and how PwC engages with the client currently.

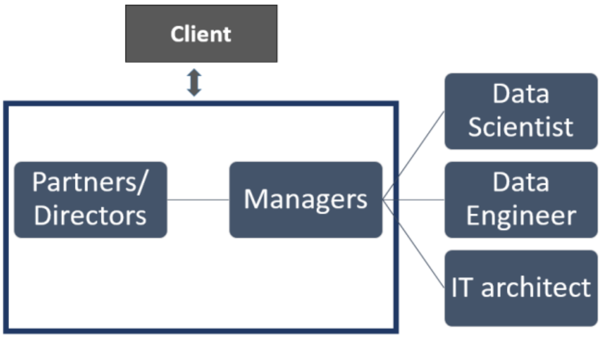


Figure 21 - Identification of actors in an AI development with respect to PwC current engagement with clients

But as the level of automation and complexity gets higher and when such application is going to be used in crucial industrial services like Finance, Healthcare, etc., there would be more stakeholders involved and PwC have already started to work on such engagements. It was also clear while speaking to some of the directors & managers at PwC that there would be more incoming actors involved in the development of AI with clear segregation of roles and responsibilities. Regulators, Functional AI experts, policymakers, testers, auditors and several AI-based research institutions would be involved. As AI spreads into more specific domains, It would not only require the knowledge and skillsets that data scientist and AI experts have but also the need for functional specialists & industry experts that could help the AI systems to produce results effectively with free of errors. Engaging the right experts and the representatives of the industry sectors would help to understand the emerging issues that AI could bring to society and can help in providing guidance to solve such issues. This can also help to support best practices in the research, development and the use of AI technology within such domains.

4.2 OVERVIEW OF THE BUSINESS PROCESS

The below figure 22 shows the actual business process flow of AI-based delivery solutions to the clients. The process flow developed was validated with some of the actors (data scientist, data analytics consultant) at PwC. To provide a standard notation of business process flow, the concept of Business Process Modeling Notation (BPMN) is used. Using White (2006) definition of BPMN as a reference, BPMN in the context of AI development can be said as:

The main objective of BPMN is to provide a notation that is readily understandable by all the business leaders right from the client that explains the core business process, to the managers and the Data Analytics team comprising of data scientist, data engineers, IT architect, and DevOps engineer who are responsible for establishing the proof of concept, gathering the data and processing the data, developing the model and implementing the model at the client space.

The main actors involved during the development of AI models would be the directors and the senior managers at the first instance who would actively engage with the client on behalf of the organization. These actors would realize the need for improvement, once the client explains their core business process. The managers would then involve the Data Analytics team and according to figure 22, Data Scientist, Data Engineer, IT architect, DevOps engineer would be represented as the Data Analytics team. The managers would mobilize the relevant actors from the Data Analytics team to develop a proof of concept which would be shared to the client. Once the client is convinced with the proof of concept, the main development and implementation would begin. The Data Analytics team would process the data and develop the model in the sequential way as mentioned below once they receive the data from the client.

⇒ Gather the data ⇒ Assess the quality of data ⇒ Process the data by cleaning, labeling, and visualizing the data ⇒ Analyze the patterns ⇒ Choose, build, and train the model with training data ⇒ Evaluate the model with test data ⇒ Deploy the model/ system at the client space ⇒ Monitor the model.

Additionally, a high-level process flow involving the main actors working on the data and building the model has been portrayed in Appendix D. It identifies the probable stages in the workflow where the dimensions of DQ and the trust factors of the model would be critical in typical development, deployment of AI models. For instance, the data engineer after receiving the data from the client has to check whether the data is valid or not and in such stages, the dimensions of DQ would come into the picture as the data engineer has to validate the data by considering the important dimension of data quality before proceeding to the next stage of the process. Similarly, once the model has been built and trained based on the training data, the data scientist has to consider factors like the reliability of the model, the performance of the model, security of the model, etc. as trust factors before deploying the final model at the client space.

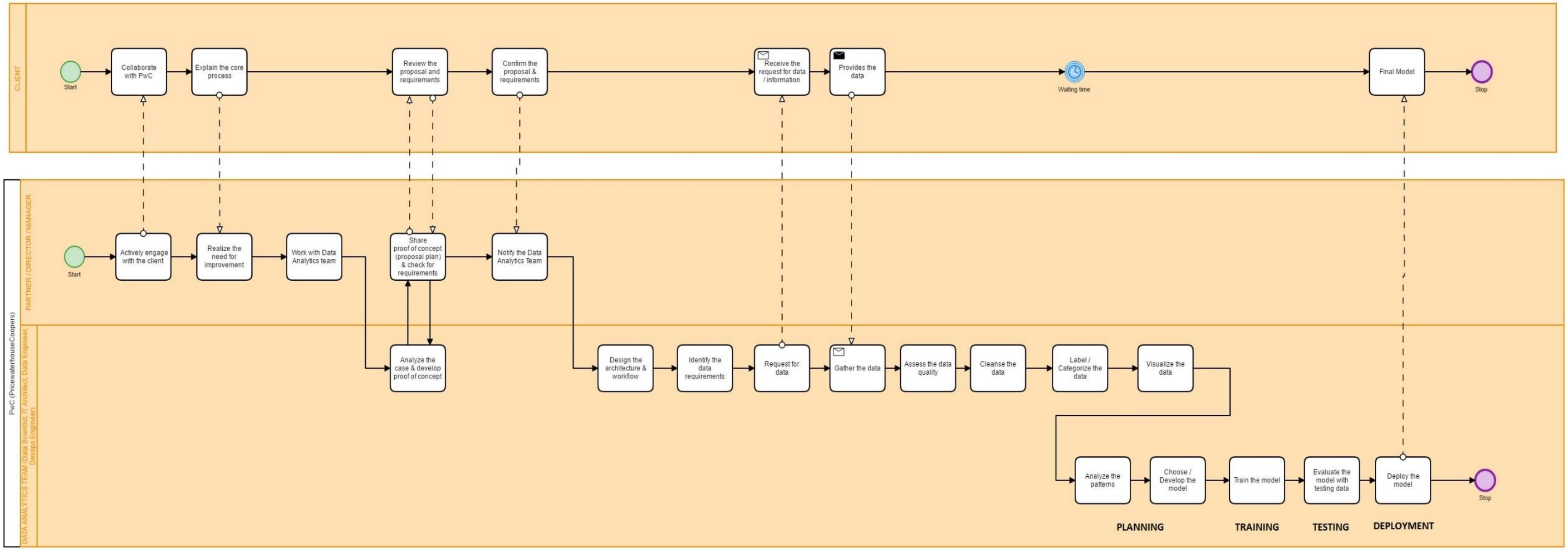

CLIENT
Start
Collaborate with PwC
Explain the core process
Review the proposal and requirements
Confirm the proposal & requirements
Receive the request for data / information
Provides the data
Waiting time
Final Model
Stop
PwC (PricewaterhouseCoopers)
PARTNER / DIRECTOR / MANAGER
Start
Actively engage with the client
Realize the need for improvement
Work with Data Analytics team
Share proof of concept (proposal plan) & check for requirements
Notify the Data Analytics Team
DATA ANALYTICS TEAM (Data Scientist, IT Architect, Data Engineer, Devops Engineer)
Analyze the case & develop proof of concept
Design the architecture & workflow
Identify the data requirements
Request for data
Gather the data
Assess the data quality
Cleanse the data
Label / Categorize the data
Visualize the data
Analyze the patterns
Choose / Develop the model
Train the model
Evaluate the model with testing data
Deploy the model
Stop
PLANNING
TRAINING
TESTING
DEPLOYMENT

Figure 22 - Simplified Business process flow - AI development

4.3 HOW ARTIFICIAL INTELLIGENCE WORKS

This section would give a snapshot of how AI works and for this demonstration, one of the subsets of AI, Machine Learning (ML) will be explained. This would give a better understanding to the readers as the research would try to emphasize the value of data, complexity of the model and the importance of trust towards such models. It is to be noted that the section won't cover the detailed or technical aspects of the model.

> Machine learning is a ***"field of study that gives the computer the ability to learn without being explicitly programmed"*** (Awad & Khanna, 2015, p. 1).
>
> ***Arthur Samuel***, one of the pioneers of machine learning

Machine learning is the process of enabling the machines to learn from data without explicitly programming it with rules as it can learn from the data provided. It is about applying the algorithm to fit the model to the data (IBM, 2018). Some of the examples of machine learning algorithm are decision trees, artificial neural networks, deep learning, Bayesian networks, and reinforcement learning. (Guo, 2017) identified seven major steps to build an efficient ML and the steps include Data Gathering, Data Preparation, Model Selection, Model Training, Model Evaluation, Hyper-parameter tuning, and Prediction.

Data Gathering - One of the most important stage as the quality and the quantity of data that is being gathered could directly determine how good the model would be.

Data Preparation - the second most important stage where the data needs to be processed (Cleaning, filtering, removing duplicates, normalization, data type conversions, etc.) and loaded into a separate place in order to prepare them for the machine learning training. Split the data into training data and evaluation data sets.

Model Selection - Choosing the right algorithm based on the context and the data.

Model Training - Using the data incrementally to improve the model's ability to make correct predictions.

Model Evaluation - Testing the model with evaluation data (data that has never been used for training). Precisely, this would show how the model would perform in the real world context considering the evaluation data to be representative of real-world context.

Hyper-parameter Tuning - Based on the results, tuning the parameters of the model in order to improve the performance.

Prediction - is where the final value of machine learning is realized. New data sets like test data can be used to check the performance of the model.

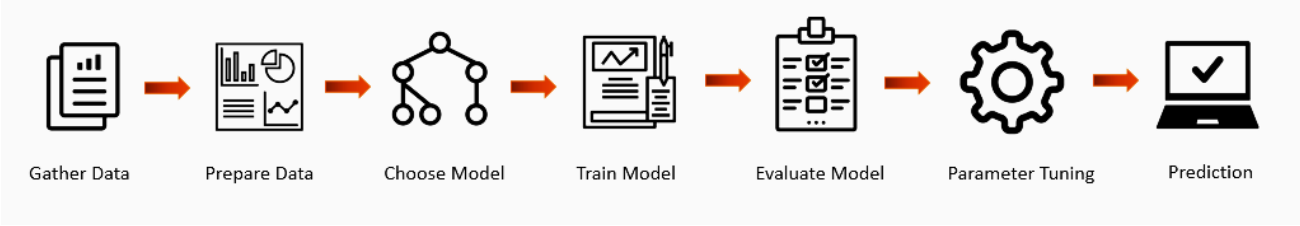
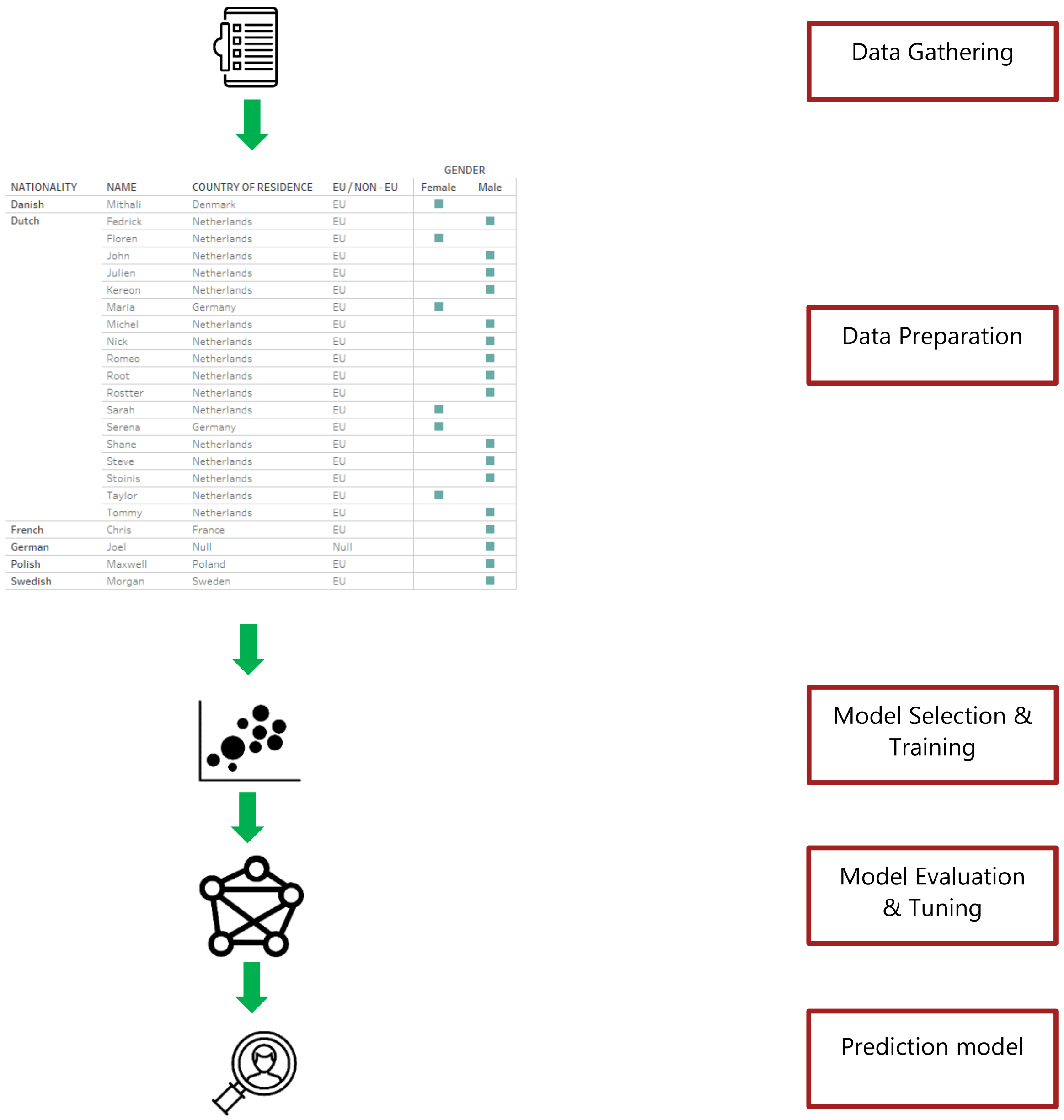


Figure 23 - Steps involved in building Machine learning

A simplified framework was proposed by (Mayo, 2018) that included five stages and it covers all the main stages as proposed by (Guo, 2017). The five stages are Data collection and preparation, Feature Selection and Feature Engineering, Choosing the machine learning algorithm and training the model, Evaluating the model, and Model tweaking, regularization, and hyper-parameter tuning. Of all these stages, only feature selection and engineering stand out as a unique stage when compared with stages proposed by (Guo, 2017).

A sample & a simplified use case has been created to give a clear picture of how ML would work in reality and use case description is as follows: A small-medium sized tech company ABC having an employee count of 25 people are planning to adopt an AI-based recruiting tool as a part of screening the candidate resumes. The company is testing by loading current employees resume (25 resumes) as an input to the model. The below use case is just to give some clarity to the readers through visual representation and it won't be emphasized with details of how the model works in the background.



				GENDER	
NATIONALITY	NAME	COUNTRY OF RESIDENCE	EU / NON - EU	Female	Male
Danish	Mithali	Denmark	EU	■	
Dutch	Fedrick	Netherlands	EU		■
	Floren	Netherlands	EU	■	
	John	Netherlands	EU		■
	Julien	Netherlands	EU		■
	Kereon	Netherlands	EU		■
	Maria	Germany	EU	■	
	Michel	Netherlands	EU		■
	Nick	Netherlands	EU		■
	Romeo	Netherlands	EU		■
	Root	Netherlands	EU		■
	Rostter	Netherlands	EU		■
	Sarah	Netherlands	EU	■	
	Serena	Germany	EU	■	
	Shane	Netherlands	EU		■
	Steve	Netherlands	EU		■
	Stoinis	Netherlands	EU		■
	Taylor	Netherlands	EU	■	
	Tommy	Netherlands	EU		■
French	Chris	France	EU		■
German	Joel	Null	Null		■
Polish	Maxwell	Poland	EU		■
Swedish	Morgan	Sweden	EU		■

Figure 24 - Illustration of the working of machine learning with a sample use case

4.4 IMPLICATION OF TRUST ISSUES WITH RESPECT TO DATA AND AI/ML MODEL

One of the major challenges during the development of AI/ML is the trust issues that are encircled not only towards the final outcome of the technology but in between the process of developing the technology. Data has been the seed to the success of AI and it is of no exception that the quality of data is a basic essential property that determines the value and trust towards the data. As highlighted by (Janssen et al., 2017), DQ issue refers to the *unfulfillment of fitness for use condition by the users which may lead to poor information quality.* When such information (processed data) is loaded to the model and if the model is going to be trained based on such information, the outcome of the results from the model would be bad & inconsistent and thereby raising the bar of trust issues over the technology.

To trust the data, one has to look for all relevant dimensions of DQ and it has to be validated recursively at every stage of handling and dealing with the data. The IQ in this context is by assessing the quality of the processed data.

A bad DQ can lead to bad IQ and a good IQ can imply a good DQ but a good DQ might not lead to good IQ.

Assessing the quality of data using the dimensions is certainly not a once-off validation as there are multiple levels of dealing with the data in the light of AI. So, dimensions of DQ would need to be assessed right from gathering the data until the stage where the data is processed and ready as information. Once such data are loaded to the model, the trust needs to be shifted towards the model behavior and its outcome. In chapter 4.3, we saw several core stages involved in creating machine learning models and when such stages are considered, the question of trust comes at every stage of ML which needs to be ensured in order build a model that is trustworthy in all aspects. Figure 25 gives an overview, where the importance of DQ, IQ (processed data), and model dimensions (trust factors) have been projected and these dimensions need to be addressed in the respective stages involved in developing an AI/ML model. For instance, in the first stage of ML, is the data gathering or data acquisition as proposed by (Guo, 2017) and in such stage, dimensions of DQ would come as a major underlying trust factor when gathering data. Dimensions like accuracy, reliability, accessibility, availability, timeliness, etc. are some of the trust factors and this kind of dimensions had to be considered for assessing the quality of data as well as the quality of the processed data. Meanwhile, the notation like training data, test data, and validation data used in figure 25 are defined as follows: **Training data** - set of data that are used to train the model, **Validation data** - set of data separate from the training data that are used to validate the model during training and **Test data** - set of new data that are used to evaluate the model after the training.

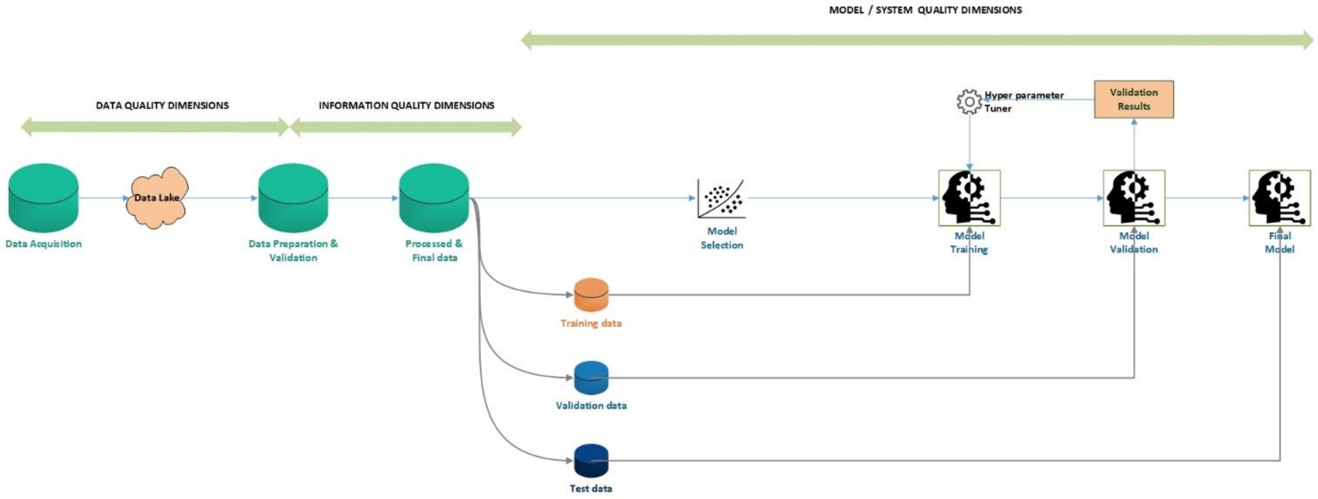
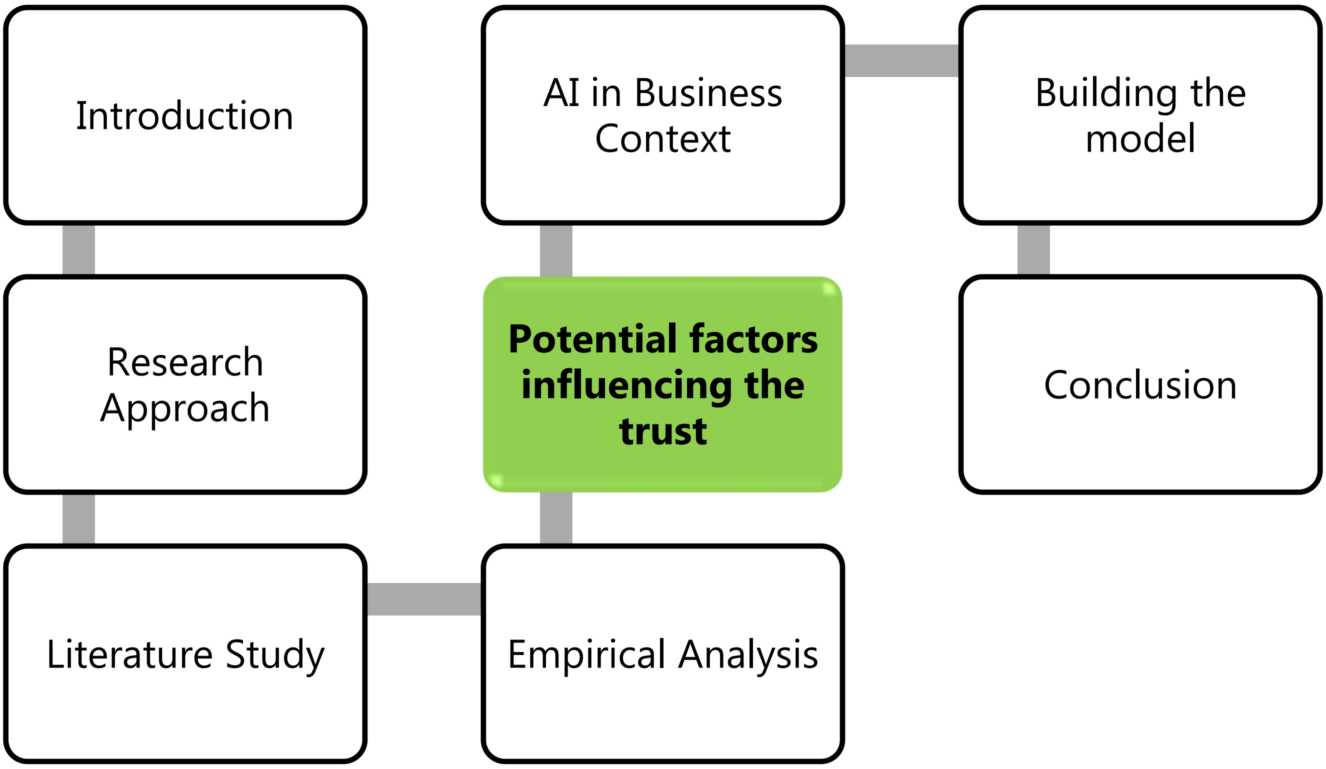


Figure 25 - Overview of trust issues with respect to data and model in the development of AI

5

POTENTIAL FACTORS INFLUENCING THE TRUST



Based on the literature study, several potential factors were identified that could influence the trust towards the AI model. Factors that could influence the trust towards the data were recognized in terms of DQ dimensions. These factors would be regarded as the independent variable and it will be further elaborated in this chapter. Only the factors that are relevant with respect to the data and model has been chosen as there are multiple factors available in the literature source especially with respect to the dimensions of DQ. There isn't single agreed solidarity on what dimensions constitutes the DQ. As a validation, It was ensured that the factors (DQ dimensions) chosen, **(1) are the most frequently recommended dimensions for DQ, (2) it is relevant to the context of AI, and (3) can highly influence the trust the technology as a whole**. Furthermore, there would be certain factors which might not contribute to DQ dimensions but those factors would be considered for placing a trust around the data process and data life cycle. A working hypothesis would be formulated for every potential factor identified from the literature review and desktop research and this hypothesis would be validated at the end of the research.

5.1 TRUST FACTORS OF DATA

ACCURACY

Accuracy can be defined in several ways, (Wang & Strong, 1996) defined accuracy as "*the degree to which the data are correct, reliable, certified and free of error*". In another definition by (Ballou & Pazer, 1985), it was specified that data are accurate only when the data values that are obtained or stored in the database correspond to the real world values. Further, (Wand & Wang, 1996) refined the accuracy notion by highlighting that the data obtained should not only be correct, unambiguous, and objective, but also meaningful and believable. In the context of AI/ML where the system would require a huge volume of data in order to learn and make decisions, accuracy would be the main preliminary factor and it needs to be ensured that the data is actually correct and is reflecting the actual state of what the user expects in terms of real-world representation during the data acquisition and processing phase. Having an accurate data upfront would really influence the outcome of the model.

Working hypothesis 1: *Having data that is **accurate** will positively influence the trust towards the data and its resulting AI model**.***

CONSISTENCY

Consistency refers to the degree to which the data is presented in a format that is same and compatible with the previous data (Wang & Strong, 1996). Consistency can refer to several aspects of the data. For instances, with respect to the value of data: the value or entries in the data should be the same in all cases; With respect to the representation of the data: the entity types and attributes should have the basic structure wherever possible (Structural consistency); With respect to the physical representation of the data: whether the physical instances of the data is in accordance with some standard format (Redman & C., 1992). In the scope of AI systems, It should also be seen whether the information (processed data) and the actual data source are still consistent and have no contradiction.

Working hypothesis 2: *Having a **consistent** data will positively influence the trust towards the data and its resulting AI model.*

COMPLETENESS

The term completeness has multiple interpretations as it can be viewed from various perspectives. At the most abstract level, Completeness can be defined as the "*degree to which the entries and main attributes are not missing from the data sets*" (Pipino et al., 2002). At a data level, it can be said as "*a set of data is said to be complete with respect to the given purpose only if the data set contains all the relevant data and mandatory attributes which shouldn't be null* (Wand & Wang, 1996). In an entity level, a data is said to be complete if the data has all the parts of the entity description (Bovee et al., 2003). (Batini et al., 2009) gave a nice illustration of completeness in the area of relational databases, where completeness is often related to the missing of null values which means a value that exists in the real world is not reflected or available in the data collection. It needs to be ensured whether all the possible states relevant to the user population are represented in the stored information (Nelson, Todd, & Wixom, 2005). To summarize the above definitions retrieved from the literature and tie them to the relevance of AI, data is said to be complete only when the main attributes & mandatory entries in the data are not null and the data is reflecting all the possible states of the user population in order to avoid biases.

Working hypothesis 3: *A data that is **complete** will positively influence the trust towards the data and its resulting AI model.*

ACCESSIBILITY

Accessibility refers to the level of difficulty in obtaining and accessing the data from the perspective of the users working on the data. Accessibility is normally tied to the openness of the data i.e. Higher the degree of data openness, higher would be the degree of accessibility with more types of data that can be obtainable (Cai & Zhu, 2015).

Working hypothesis 4: *Data that are highly **accessible** will positively influence the trust towards the data and its resulting AI model**.***

SECURITY

In the revolution of Information Technology (IT), security has always been the stepchild for IT. In the relevance of data, Security refers to the extent to which the access to data is restricted appropriately in order to maintain its security (Wang & Strong, 1996). (Miller, 1996) highlighted two aspects of data security: (1) protecting the information from people (logical security) and (2) protecting the information from natural disasters (disaster recovery planning). The data received from the third party or another external source will not be trusted and used to its full potential if the data contained in the file is highly insecure especially when the data holds sensitive information.

Working hypothesis 5: *Data that is **secured** will positively influence the trust towards the data and its resulting AI model.*

PRIVACY

When organizations start to invest in AI, they would need to be hyper-vigilant about how their data and resulting model making predictions based on the data are lined up with new regulations and compliances especially with EU General Data Protection Regulation (GDPR) coming into the picture. Users would expect the organization that their personal data at any cost is not being compromised or used by the machines or the individuals in the organization unlawfully and ensuring that they are in compliance with the data protection laws and policies.

Working hypothesis 6: *Respecting and **securing** the **privacy** of the data will positively influence the trust towards the data and the resulting AI model.*

RELEVANCY

One of the key components of DQ is relevancy as every user like data scientist, data engineers working on the data see whether the data obtained is actually addressing the needs they expect in order to feed the data to the model. If not, the users would find that the information obtained is inadequate & it doesn't matter how well the data relates to other sets of dimensions as relevancy is one of the prime dimensions of DQ. In the light of obtaining and processing the data with respect to AI, relevancy can be termed as "*the extent to which data obtained is relevant to the specific domain and purpose of interest in a given context"* (Bovee et al., 2003). It also indicates whether the data contains the required variable in the right form and whether the data are drawn from the population of interest (Kenett & Shmueli, 2016). Relevancy would be more of an evaluating factor once the data is received and processed by cleaning, filtering, labeling and standardizing the data to ensure whether the processed data is making the actual relevance to the context.

Working hypothesis 7: *Data that is* ***relevant*** *to the purpose of the goal will positively influence the trust towards the data and its resulting AI model****.***

TIMELINESS

Timeliness is a very crucial factor as we are in the age of big data where there is a high probability that the content of the data changes very frequently. Timeliness refers to the degree to which the data or information is up to date or more precisely, it refers to the degree to which the data obtained reflects the current state of the world that represents (Nelson et al., 2005). In most of the literature, timeliness was referred to as currency. The definition of timeliness proposed by (Jarke, Lenzerini, Vassiliou, & Vassiliadis, 2003) would be in sync with the context of the data in AI. The researcher refers to timeliness in terms of currency and volatility. Currency describes when the data has entered into the sources and Volatility describes the time period for which the data is valid in the real world.

Working hypothesis 8: Timeliness *of the data will positively influence the trust towards the data and its resulting AI model.*

INTERPRETABILITY

Interpretability refers to the extent to which the data obtained is in proper language, symbols, units, etc. with clear definitions (Pipino et al., 2002). To satisfy the constructs of interpretability, the data should be both intelligible and meaningful. Intelligent data is capable of being understood by the user and meaningful data conveys to the user with some sense, significance or meaning (Bovee et al., 2003). If the data is either unintelligible or meaningless, all the other dimensions considered would be irrelevant and if such data are fed to the model, it can possibly lead to bad outcomes and it would be tedious for the users to understand the background behind such outcomes.

Working hypothesis 9: *Data that is* ***interpretable*** *will positively influence the trust towards the data and its resulting AI model.*

RELIABILITY

Reliability refers to the extent to which the data obtained is actually correct and reliable (Wang & Strong, 1996). It also indicates whether the data obtained from a source can be counted on to convey the right information which can be viewed as the correctness of the data (Batini & Scannapieco, 2016). It was also specified that reliability can be closely linked to the probability of preventing errors or failures, consistency and dependability of the output information, interpreting as a measure of agreement between the expectations and capability (Wand & Wang, 1996). One common interpretation that can be extracted from literature in the context of AI is that reliability should able to indicate whether the data that has been obtained from an external party or any sources can be counted on to trust the data, conveying the right information, etc.

Working hypothesis 10: *Data that is* ***reliable*** *will positively influence the trust towards the data and its resulting AI model.*

AVAILABILITY

Availability of data is the extent to which the data is present, obtainable, and ready for use (Batini & Scannapieco, 2016). Availability and accessibility of data are dependent on each other. If the data that is required is physically available, it should be ensured that such data are accessible as well. As (Cai & Zhu, 2015) highlighted that higher the degree of openness of the data, higher would be the degree of availability and accessibility.

Working hypothesis 11: *Data that is highly* ***available*** *will positively influence the trust towards the data and its resulting AI model.*

AUDITABILITY

In general, Auditability can be termed as the process of standardized evaluation and examination of records, statements, process & controls of an organization to see how far the financial & non-financial statements projects a view of the actual reality. In the context of the data, auditability refers to the extent to which auditors can fairly evaluate the accuracy and integrity of the data within the rational time and limits of the manpower during the entire process of the data life cycle (Cai & Zhu, 2015). These data life cycle in the light of AI includes right from gathering the data to feeding the data to the AI model. At every phase, there should be a need for audits to examine and evaluate every change made with respect to the data. So, every changes & assumption made with respect to the data to be documented in order to examine the data process and handling. This could improve the integrity and trust towards the data.

Working hypothesis 12: Auditing *the data & its process will positively influence the trust towards the data and its resulting AI model.*

BIAS-FREE

Most of the academic literature has used the term ‘Objectivity’ in place of the word ‘bias’ as a dimension of DQ as objectivity also focuses on the importance of having unbiased and impartial data. Objectivity refers to the extent to which the data is unbiased, unprejudiced, and impartial (Wang & Strong, 1996). Precisely in the frame of AI, bias refers to the inclination of prejudice towards or against a person, object, or position which can arise in many ways in the state of AI system and these kinds of biases mainly originates from the data that is being used to train the AI model (European Commission, 2019). In the data-driven AI system, there is a high possibility that the AI system can demonstrate bias if there is bias in the data collection and training. In certain cases, bias can result in discriminatory outcomes in terms of gender, race, ethnicity, religion, etc. due to data that is being fed to the system favoring the inclination towards a particular subject or family.

BIAS FREE = OBJECTIVITY

Working hypothesis 13: *Having data that is* ***bias-free*** *will positively influence the trust towards the data and its resulting AI model.*

USEFULNESS

Usefulness refers to the extent to which the information is applicable and helpful for the task at hand (Wang & Strong, 1996). It indicates whether the data that is obtained is useful by means of reliability of data at the first instance, the objectivity of the data (unbiased and impartial), the relevancy of the data with respect to the population of the interest, consistency, and completeness of the data. The dimensions like reliability, objectivity (bias-free), relevancy, consistency, completeness, and interpretability of the data should be validated first in order to ensure the overall usefulness of the data. In light of big data & AI, these dimensions should act as a prerequisite to confirm the usefulness of the data.

Working hypothesis 14: *Data that is* ***useful*** *will positively influence the trust towards the data and its resulting AI model.*

Briefly, to determine the quality of data and to place trust in the data, we must ensure that:

1. Data received is accurate, free from errors, reflecting the actual state of context that is expected. **(Accuracy)**

2. Values in the data are the same, consistent and are in accordance with standard format and have no contradiction between the actual data source and processed data. **(Consistency)**

3. The main attributes & mandatory entries in the data are not null and the data is reflecting all the possible states of the user population in order to avoid biases. **(Completeness)**

4. There is a high degree of openness to data that is readily available and accessible especially from the genuine source. **(Availability & Accessibility)**

5. The data obtained is relevant to the specific domain and purpose of interest in a given context and having the relevant variable in the right form. **(Relevancy)**

6. The data is highly safeguarded especially when the data contains personal and sensitive information. **(Security)**

7. Understanding & find the meaning of the data and the language, metrics & symbols used is understandable to the users working on the data. **(Interpretability)**

8. The validity of the data in the real world and having up to date information. **(Timeliness)**

9. Data & contents of the data are unbiased, impartial in order to avoid the inclination towards a particular subject. **(Bias-Free)**

10. The changes made in the data are documented, understand the responsibility of data owners and the process of data handling in order to audit the data. **(Auditability)**

11. The data obtained from a source can be counted on to convey the right information. **(Reliability)**

12. Meeting the compliance standards with respect to EU's GDPR and other data protection policies. **(Privacy)**

13. The resultant data is useful by validating the dimensions of reliability, relevance, bias-free, consistency, completeness, etc.. **(Usefulness)**

5.2 TRUST FACTORS OF THE MODEL

RELIABILITY

Reliability refers to the system’s ability to work appropriately with a range of inputs and in range of situations in order to scrutinize the AI system and to prevent accidental harms (European Commission, 2019). It is critical that the results of the AI system results are reliable and perform as desired in order to place trust in AI systems. Reliability can be in terms of the outcome of the final model, selecting the appropriate model/algorithm, etc.

Working hypothesis 1: *AI systems that are* ***reliable*** *will positively influence the trust in AI.*

ACCURACY

Accuracy refers to a system ability (AI system) to make the correct judgment, predictions, recommendations, or decisions based on the data and the model. A well-formed development and evaluation process can support and mitigate the unintended risk from inaccurate results (European Commission, 2019). Users and business leaders would expect a high level of accuracy especially in a very critical environment that directly affects the human lives (Healthcare, Financial sectors, etc.).

Working hypothesis 2: Accurate *results produced by an AI system will positively influence the trust in AI.*

AUDITABILITY

The ability of a system that undergoes the assessment on its algorithms, data and design process is called as auditability (European Commission, 2019). It is necessary for the auditors to evaluate and assess the entire life cycle of the AI system and in order to validate such system, they would expect the system to have traceability and logging mechanism right from the early design phase of the AI system till the deployment of the system as that could help in empowering the auditors to asses such systems and improve the trust of technology. Factors like traceability would go hand in hand with auditability as traceability refers to the capability of keeping track of the systems data, development, deployment, process, etc. through a documented recorded identification. Assessing such documents would improve the auditability and trust in AI.

Working hypothesis 3: Auditing *the AI system will positively influence the trust in AI.*

BIAS-FREE

As mentioned in chapter 5.1, where bias-free is considered as one of the trust factors towards the data, It would also require high attention with respect to the model as well. Bias refers to the inclination of prejudice towards or against a person, object, or position which can arise in many ways in the state of AI system (European Commission, 2019). The bias that exists in the data can directly influence the model to be biased. Having a model that needs to bias-free has to be ensured right from the data that is being collected, processed, trained, etc. to the outcome of the model.

Working hypothesis 4: *AI systems that produce results without* ***biases*** *will positively influence the trust in AI.*

CONSISTENCY

Consistency refers to the ability of the AI systems that can exhibit the same behavior without any ambiguity when repeated under the same conditions. The term consistency is closely related to the factor "Reproducibility" as recommended by (European Commission, 2019) which is considered as one of the factors to influence the trust towards AI.

CONSISTENCY = REPRODUCIBILITY

Working hypothesis 5: *AI systems that are* ***consistent*** *with the results, predictions, recommendation, etc. will positively influence the trust in AI.*

ETHICAL

Ethics is considered a sub-field of philosophy in an academic discipline. Normally, Ethics deals with questions like "What is considered to be a good action?", "What is justice?", "What is the value of human life?" etc. In the academic discipline, ethics can be classified into four major types of research: Meta-ethics, normative ethics, descriptive ethics, and applied ethics. In the context of AI ethics, it is generally viewed as an example of applied ethics and would focus on normative issues raised during the various stages [Design, Development, Implementation] of AI (European Commission, 2019). Applied ethics generally deals with: What we are permitted to do in a specific situation or a particular domain of (unprecedented) possibilities for action. It needs to be ensured that AI is in compliance with ethical norms & principles, related core values, and fundamental rights.

Working hypothesis 6: *Having an AI system that is* ***ethical*** *will positively influence the trust in AI.*

GOVERNANCE

The term governance in the context of AI aims to close the gap that exists between accountability and ethics in technological development. The concern on the governance of AI is getting high attention as the technology has already started creating impacts in many industrial sectors. The governance in AI involves: identifying the answers to the questions surrounding the safety of AI, what legal and institutional sectors to be involved, who has the control and access to personal data, what are the roles of moral & ethical institutions when interacting with AI, control, and monitoring of the algorithm and the data (Rouse, 2018). In the light of AI with respect to the data, a separate governance has to be established that should involve: quality and integrity of the data, its relevance in the light of the domain in which the AI system is used, data access, protocols, ability to process data that consider data as an asset and protect privacy (European Commission, 2019). Governance would actually serve as an umbrella for most of the influential factors (Privacy, Security, ethics, accountability, auditability, compliance, etc.).

Working hypothesis 7: *Establishing the* ***governance*** *around the model will positively influence the trust in AI.*

PERFORMANCE

In order to validate & improve the performance of the AI system, stakeholders would expect that the outcomes of the AI model are aligned with their expectation and performs at the desired level of precision and accuracy with unbiased results. Furthermore, the users of the AI systems would solely rely on the performance of the system in critical sectors like healthcare, financial services, etc. where the users could make a decision based on the prediction or recommendation made by the system.

Working hypothesis 8: *Increase in the* ***performance*** *of the AI system will positively influence the trust in AI.*

PRIVACY

When AI systems like machine learning are used, one would expect the system to guarantee its privacy and data protection throughout a system's entire lifecycle. Privacy is not only tied to the data but also to the final model which uses the data as the main source in order to predict the outcomes. For instance, the output that the AI system generates and gives a personalized recommendation to each user is mainly based on users historical data and in such case, the users would expect that their data is not being used unlawfully by the model and within the data gathering and processing phase, whether their data is properly protected, understanding how their data is being processed by the model.

Working hypothesis 9: *AI systems that respects and* ***safeguards*** *the user* ***privacy*** *will positively influence the trust in AI.*

SECURITY

The data that is used by AI system components and the algorithm itself should be secured from unauthorized access and adversarial attack. Also, the training and test data should be secured. (Siau & Wang, 2018) highlighted that operational safety around the AI system and data security would be one of the important factors that can influence the trust in AI as people are unlikely to trust anything that is too risky to operate. Data Security, for instance, would be critical as systems like Machine learning (ML) relies on a large amount of data, making the security of data a concern.

Working hypothesis 10: *AI systems that are* ***secured*** *will positively influence the trust in AI.*

TRANSPARENCY

External users of AI systems, regulators, auditors expect the model to be transparent. Transparency is concerned with conveying the structural details of the model, descriptive properties of the training data, and any evaluation metrics from which the likely behavior of the model can be inferred. To trust such AI systems, one would expect to understand on how the AI models had been programmed and what function will the model perform in certain conditions that can help to shed light on the black box models. Transparency is considered as one of the dimensions of interpretability where the transparency helps to answer the question: How does the system work? (Oxborough, Rao, Cameron, & Westermann, 2018).

Working hypothesis 11: *Having an AI system that is* ***transparent*** *will positively influence the trust in AI.*

EXPLAINABILITY

As the AI system gets more complicated, there would be more and more decision making being performed by the algorithmic black box. In order to have confidence in the outcomes made by such AI systems and to cement the trust of the stakeholder as they want to know why specific outcomes occur, it would be necessary to know the rationale of how the algorithm arrived at its decision or recommendation. If such AI systems can be made explainable, it would give more space for the users to anticipate how the system would behave or predict and ideally making users interpret the behavior and probable outcomes (Interpretability). Explainability is considered to be one of the core critical factors that influence most of the other factors encircled around AI in order to trust the technology. It is one of the powerful tools that AI can have which can help to detect the flaws, biases in the data, helping users to interpret the model behavior, maximizing the performance based on the potential weaknesses identified from the model, etc. (Oxborough et al., 2018).

Working hypothesis 12: *AI systems that can* ***explain*** *the decisions made will positively influence the trust in AI.*

INTERPRETABILITY

It is assumed that most of the AI systems used for automated decision making are based of ML over big data where the algorithm maps the user features into a class that can predict the behavioral traits of the individual such as credit risk, health status, etc. without exposing the reasons behind the predictions which we term as a "Black box AI system" (Guidotti, Monreale, & Pedreschi, 2019). To address this problem and & to improve the trust over such systems, it was recommended to design interpretable models and have the ability for the machine to explain its conclusion or actions (Siau & Wang, 2018). This could ideally help the users of such systems to understand the rationale for the outcomes and the process of deriving the results. (Oxborough et al., 2018) classified the two dimensions of interpretability as transparency and Explainability.

Working hypothesis 13: Interpretable *AI models will positively influence the trust in AI.*

ACCOUNTABILITY

As there is an increasing awareness on the responsible approach to AI in order to ensure safe, beneficial, and fair use of AI systems, (Dignum, 2018) proposed the principle of accountability, responsibility, and transparency (ART) where accountability was defined as the "the ability to explain and try justifying one's decision & actions to the business leaders, users, indirect stakeholders with whom the system interacts and influences". As a prerequisite for assessing accountability in AI, it requires both the function of guiding action and the function of explanation (making the decision in a broader context and classifying them along with the moral values).

Working hypothesis 14: *Establishing the* ***accountability*** *principle will positively influence the trust in AI.*

USEFULNESS

A model is said to be useful only when AI models spit accurate results, showing consistency and reliability with the outcomes, making the AI system explainable & transparent that helps the users, auditors to interpret and audit the model behaviors, securing the data, ensuring compliance and regulation and finally establishing the governance around the model and the data.

Working hypothesis 15: *AI systems that are* ***useful*** *will positively influence the trust in AI.*

5.3 CONCEPTUAL MODEL

Upon identifying the potential dimensions of DQ and the potential factors of the AI model from the literature that will positively influence the trust in AI, a conceptual model has been drafted which includes potential factors towards the data and model influencing the trust in AI. This conceptual model will serve as a guide for the rest of research and for the interviews and case studies in order to understand the importance of these factors from the perspective of the relevant actors who are involved in the development of AI and what factors they would see as important in order to trust the data and the system. In the end, it could also introduce an additional set of trust factors & the possibility of limiting the trust factors by setting them a priority and it will be made to ensure that those factors are also reflected in the final version of the trusted AI model.

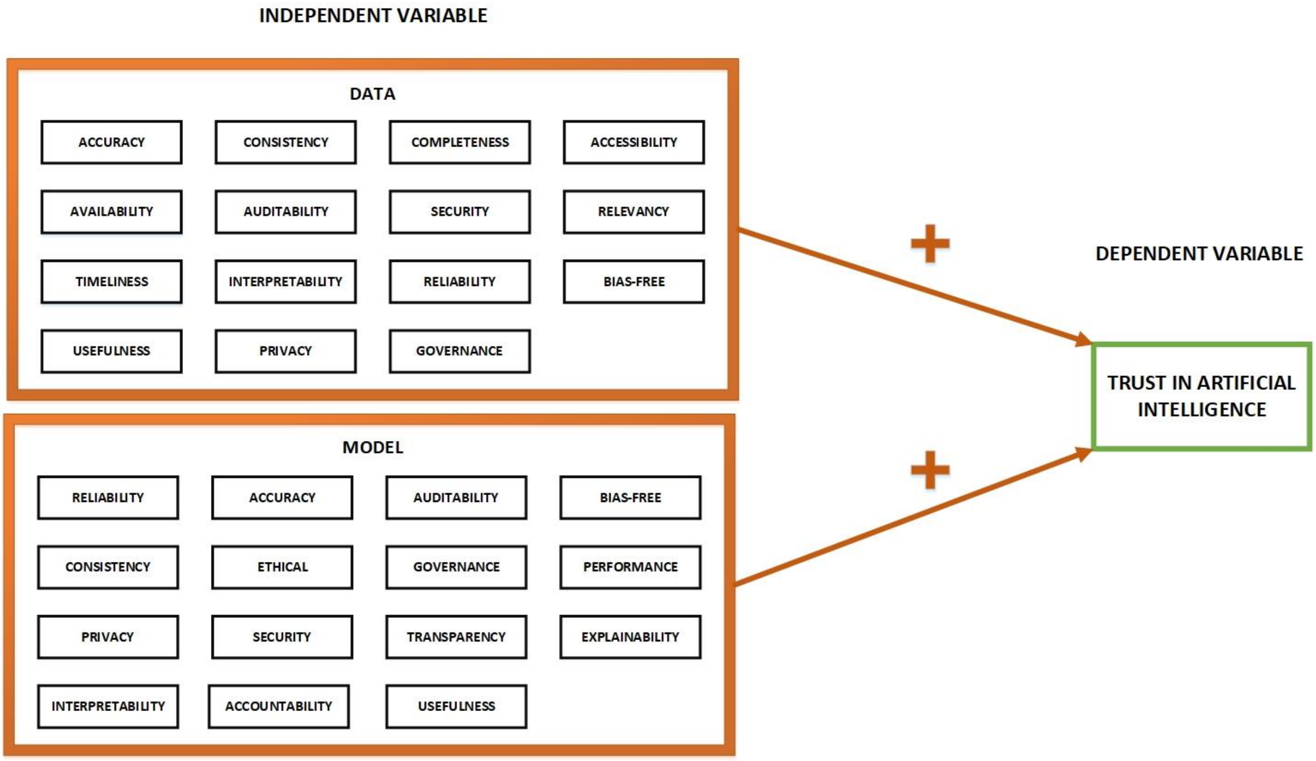
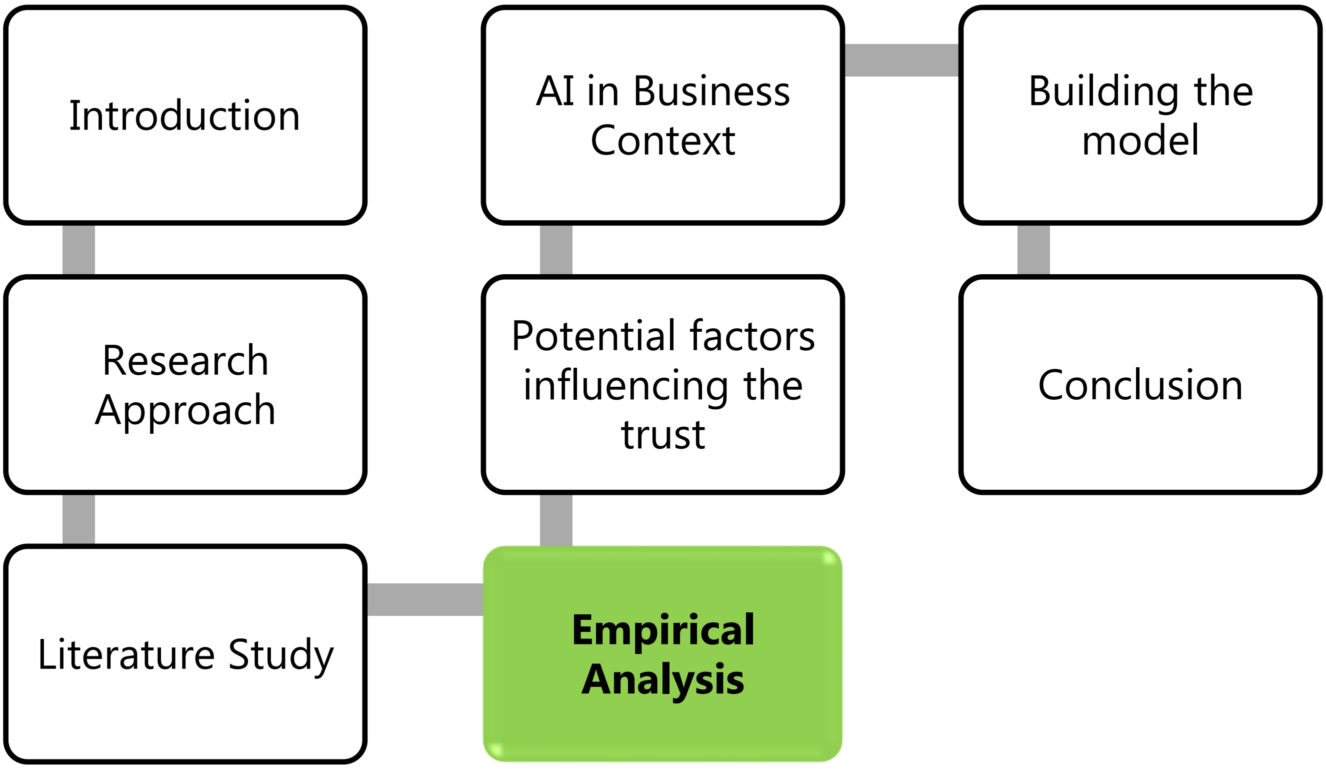


Figure 26 - Conceptual model

6

EMPIRICAL ANALYSIS



6.1 INTERVIEW PROCESS

A structured interview approach was followed in this research as the main goal is to understand what factors and dimensions of trust influence the technology from the perspective of the actors (Directors, managers, data scientist, data engineer, AI experts & specialists, risk advisor, and auditors) involved in the development of AI. Several conditions were used for scoping the sample from the population. For this research, **20 people** from PwC were initially invited to participate and out of which **16 people** had agreed to participate (initial phase of the research). It was made sure that the selected participants fall into any of the below criteria

1. The respondent is aware of the technology, its value & associated risks

2. The respondent has experience in working on AI and automation based projects.

3. The respondent has sufficient experience in working on data.

4. The respondent should either be a data scientist, data engineer, developer, AI expert, Partner or Directors at management level, Auditors, Advisors / Consultants.

To encourage a diverse set of opinions and gain new insights, respondents from different departments, designation & roles (Advisory, Consulting, Assurance team, etc.) were approached and to increase the research validity,

- Respondents with the positions of data scientist, data engineer & stewards, etc. were interviewed to understand the value of data, challenges and its influence on the AI model.

- Respondents proficient in dealing and mitigating the risk of emerging technologies were involved (risk assurance & advisors).

- Finally, the AI experts & Specialist from PwC were also involved in this study.

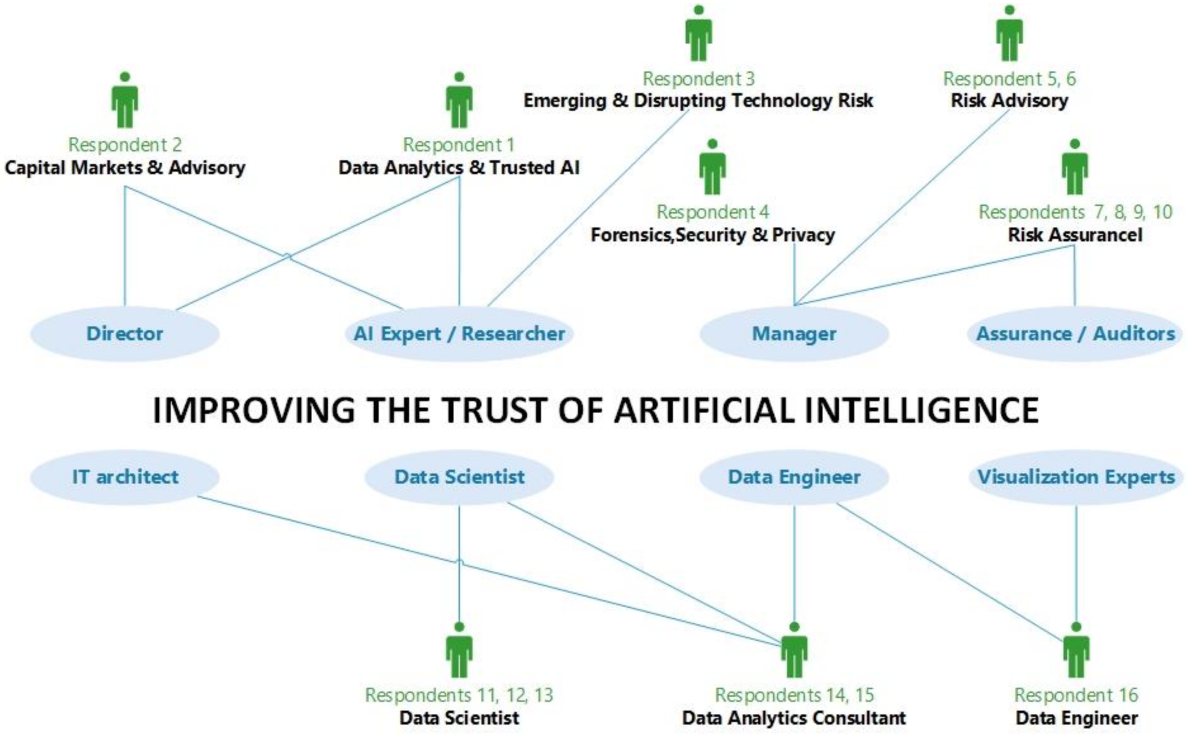


Figure 27 - Overview of the interview respondents

To avoid biases in the interviewee response and to improve the clarity of the questions & overall interview process, the below techniques were followed as recommended by (Sekaran & Bougie, 2016),

Funneling – Asking some open-ended question to the respondents to give some impression about the situation.

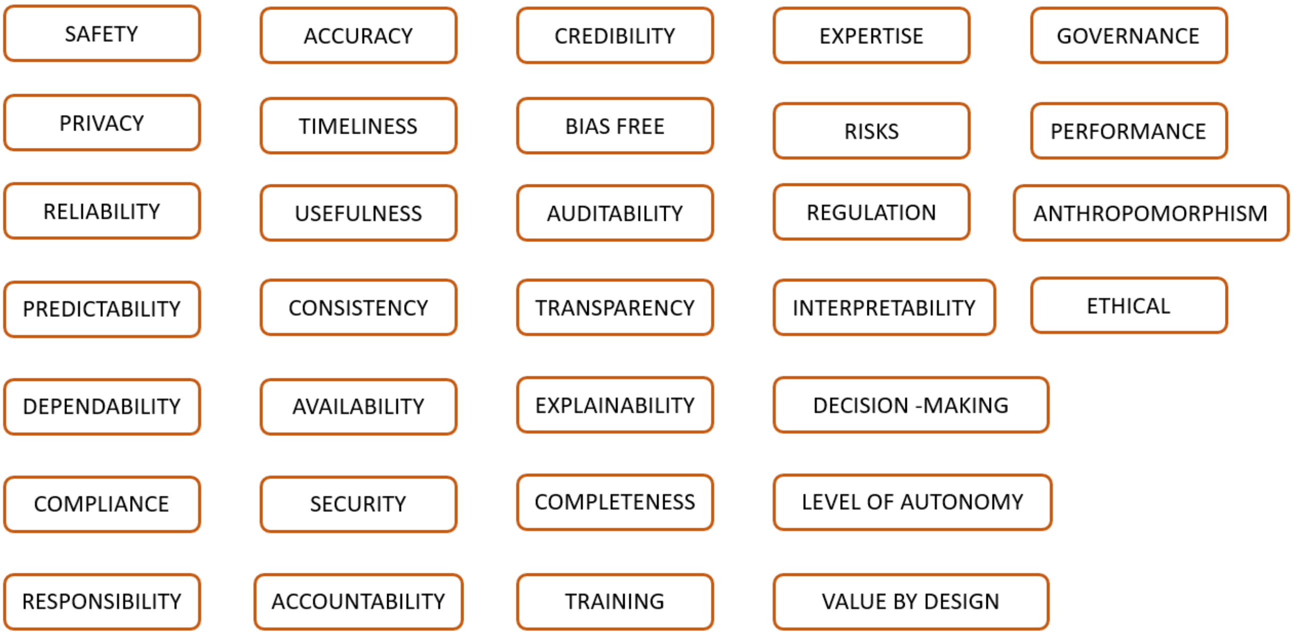
Unbiased Question – Ensuring that unbiased questions are asked to minimize the bias in the responses.

Clarifying Issues – It is advisable to restate or rephrase the information provided by the respondent to make sure that the information provided by the respondent is well understood and is in line with the context of the research.

Helping respondents to think through the issues – Rephrasing the questions in a simple way and providing sample use cases in order to help the respondent understand the issue and question posed.

Notes making & Interview recording – It is believed that information recalled from the memory is imprecise and often incorrect and it would introduce more bias into the research. So, keynotes were taken and interviews conducted were recorded post getting the approval consent from the respondents. The anonymity and privacy of the respondents will be respected and safeguarded throughout the entire research.

Initially, a list of factors that were identified, studied from the scientific literature and other primary sources (business article) were presented to the respondents to make them familiar with the context of the issue. Furthermore, some potential use case examples were explained upfront to accentuate the importance of trust and factors influencing the trust towards the AI. The respondents can then identify the factors from the visual aids (Slide) that was presented upfront in order to understand the important factors from their perspective. The respondents were expected to make a clear explanation when they identify the important factors from their perspective as there could be factors which could be tied to both the data and the model. For instance, accuracy is considered as one of the main dimensions of DQ and it can also be tied to the model as a factor in terms of accuracy of the results or the model itself.



SAFETY	ACCURACY	CREDIBILITY	EXPERTISE	GOVERNANCE
PRIVACY	TIMELINESS	BIAS FREE	RISKS	PERFORMANCE
RELIABILITY	USEFULNESS	AUDITABILITY	REGULATION	ANTHROPOMORPHISM
PREDICTABILITY	CONSISTENCY	TRANSPARENCY	INTERPRETABILITY	ETHICAL
DEPENDABILITY	AVAILABILITY	EXPLAINABILITY	DECISION -MAKING	
COMPLIANCE	SECURITY	COMPLETENESS	LEVEL OF AUTONOMY	
RESPONSIBILITY	ACCOUNTABILITY	TRAINING	VALUE BY DESIGN	

Figure 28 - Presentation of trust factors to the respondents

6.2 INTERVIEW ANALYSIS & FINDINGS

All the interviews that were conducted were recorded upon the approval from the respondents and each interview recordings were transcribed manually to text for analyzing the transcripts. Though there are applications that can easily convert the recorded audio conversation to text, old & traditional method of transcribing them manually was adopted in order to respect the privacy of respondents & protect the confidential information. To analyze the transcripts in detail and to derive the most important factors especially towards the data and the model, two approaches were used,

1. Creating a summary of every interview transcripts in a form pictorial representation

2. Doing a content analysis by coding and analyzing the data using NVivo qualitative software

First, for every interview transcripts, a corresponding summary was created in a form of pictorial representation as that could help the readers to perceive the information easily from that summary snippet of interviews. The summary snippet of all the interviews could be found in Appendix A. Figure 29 is a snapshot from one of the interviews conducted with a Data Scientist at PwC. When asked about the most important factor towards the trust in AI, the data scientist highlighted that they would normally pay attention to the data as that influences the resulting AI model and would see factors like the ***transparency*** of the data itself, ***completeness*** of the data as they think in terms of the sparsity of the data and the ***consistency*** of the data. The data scientist emphasized the importance of having consistent data by giving a clear interpretation,

> *If you look at E-commerce Company where you track all the users or the daily sales, you expect the data to be consistent as the model is going to be built based on those data.* ***The more inconsistent the data, the more you lose the trust towards the outcome of the model****.*

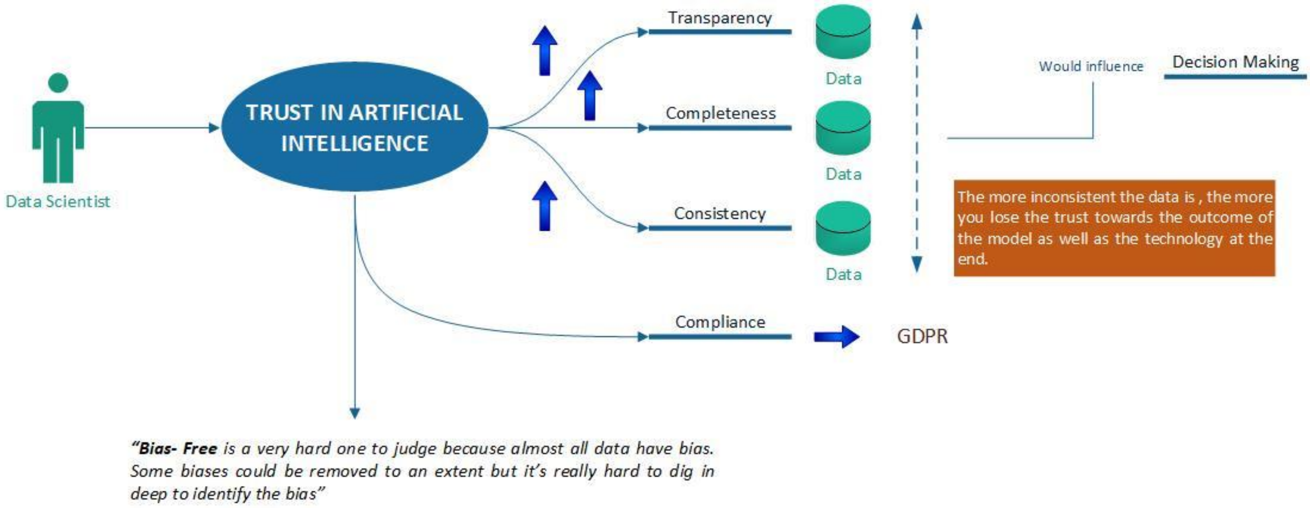

Data Scientist
TRUST IN ARTIFICIAL INTELLIGENCE
Transparency
Completeness
Consistency
Compliance
Data
Data
Data
GDPR
Would influence
Decision Making
The more inconsistent the data is , the more you lose the trust towards the outcome of the model as well as the technology at the end.
"Bias- Free is a very hard one to judge because almost all data have bias. Some biases could be removed to an extent but it's really hard to dig in deep to identify the bias"

Figure 29 - Snapshot showing the perspective of a Data Scientist towards trust in AI- Respondent 11

It was clear that most of the respondents from the perspective of a data scientist, data engineer, and data analytics consultants paid high attention to the data and how it can influence the outcome of the model. Similarly, when one of the data engineers was interviewed, the respondent cited that data has to be representative of the real world which is more towards the ***usefulness*** of the data. Having data that is ***free of bias***, the ***usefulness*** of the data, and ***accuracy*** of the data were some of the most crucial factors that data engineers would see as important to trust the data. The data engineer clearly made a distinction between the resulting model and the data, *If you want to make a distinction between what technology and data, at least in terms of technology, we don't have that many questions as it is purely about the data that we care and if* ***the data that we feed into the model is trustworthy, then the resulting model & its outcome would be efficient and safe****.*

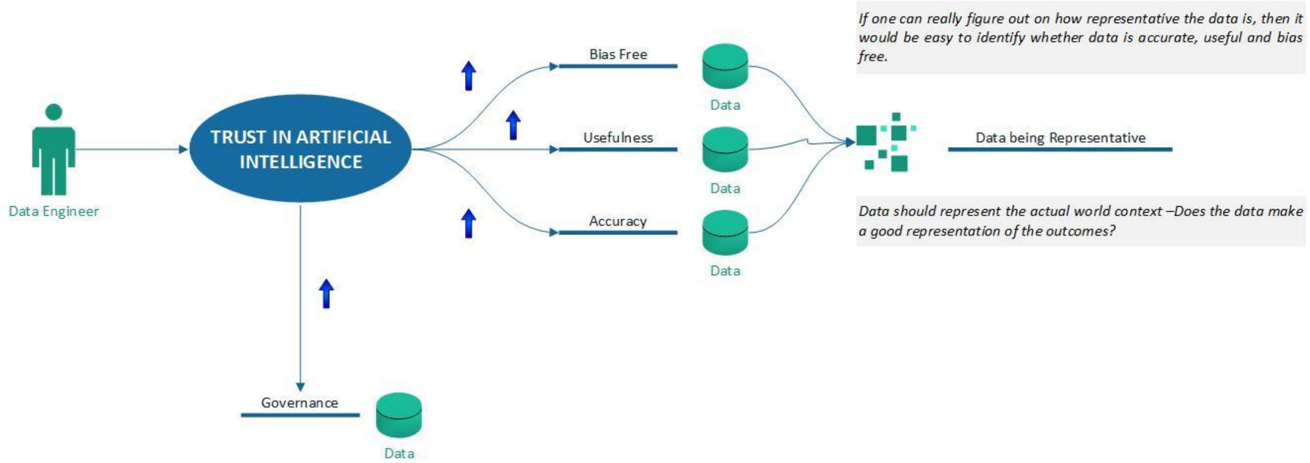

Data Engineer
TRUST IN ARTIFICIAL INTELLIGENCE
Bias Free
Usefulness
Accuracy
Data
Data
Data
Governance
Data
Data being Representative
If one can really figure out on how representative the data is, then it would be easy to identify whether data is accurate, useful and bias free.
Data should represent the actual world context –Does the data make a good representation of the outcomes?

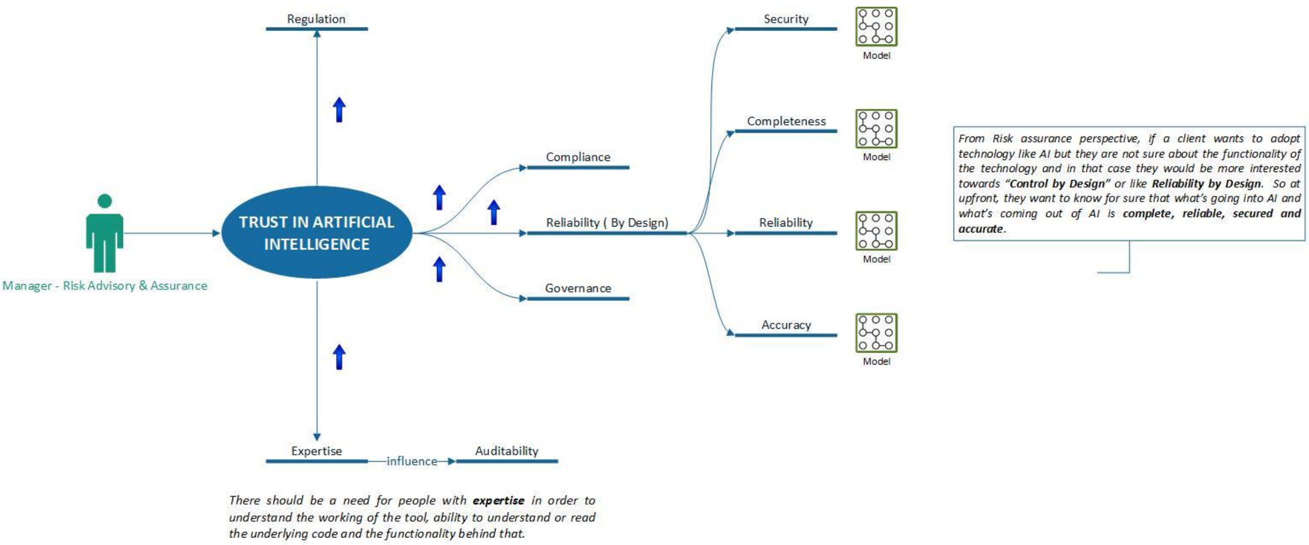
Figure 30 - Snapshot showing the perspective of a Data Engineer towards the trust in AI-Respondent 16

Additionally, It was realized the need for having robust governance set up over the data, as in reality, it hasn't been paid much attention to data governance and It would be a good practice to have a governance procedure in place in order to avoid certain risk with the data. Governance can be on a certain part of the AI systems like it can be on the data generating process or one can have governance on model building.

Deriving a summary of every interview transcripts ideally gave a direction on what factors does every actor see as important and it was quite evident that respondents coming from risk advisory & assurance background value their importance more towards the model as well as the data, while the respondents representing as a data scientist, data analytics consultants were mostly emphasizing the importance towards the data. For example, one of the managers from the risk advisory department felt that factors like ***governance, security, completeness, reliability, and accuracy*** would be the major factors that a risk advisor would see upfront and all these trust factors are tied to the model. The manager also emphasized what these trust factors would mean to them from the risk advisory perspective,

*From Risk advisory & assurance perspective, if a client approaches us and says if they want to adopt technology like AI but they are not sure about the functionality of the technology and in that case, they would be more interested towards "*__*Control by Design*__*" or like "*__*Reliability by Design*__*". So at upfront, they want to know for sure that what's going into AI and what's coming out of AI is* ***complete, reliable and accurate****.*

There was an interesting notion made by the manager on the ***governance*** as it is something that PwC would like to advise from the risk advisory perspective. When there is a tool or a system, one has to see how those tool is being governed like the changes made in the tool, who can access the tool, what are the parameters set & who can change the parameters and coding in the tool. If such governance is set, it can easily help the auditors to ***audit*** such complex systems. However, people should have ***expertise*** when such a system needs to be audited. The auditors should able to understand the working of the tool, read the underlying code and the functionality behind that in order to provide a seal of trust to the clients when they approach PwC to audit such AI systems. So expertise would be an important factor to a human actor as that can also influence & improve the trust in AI.

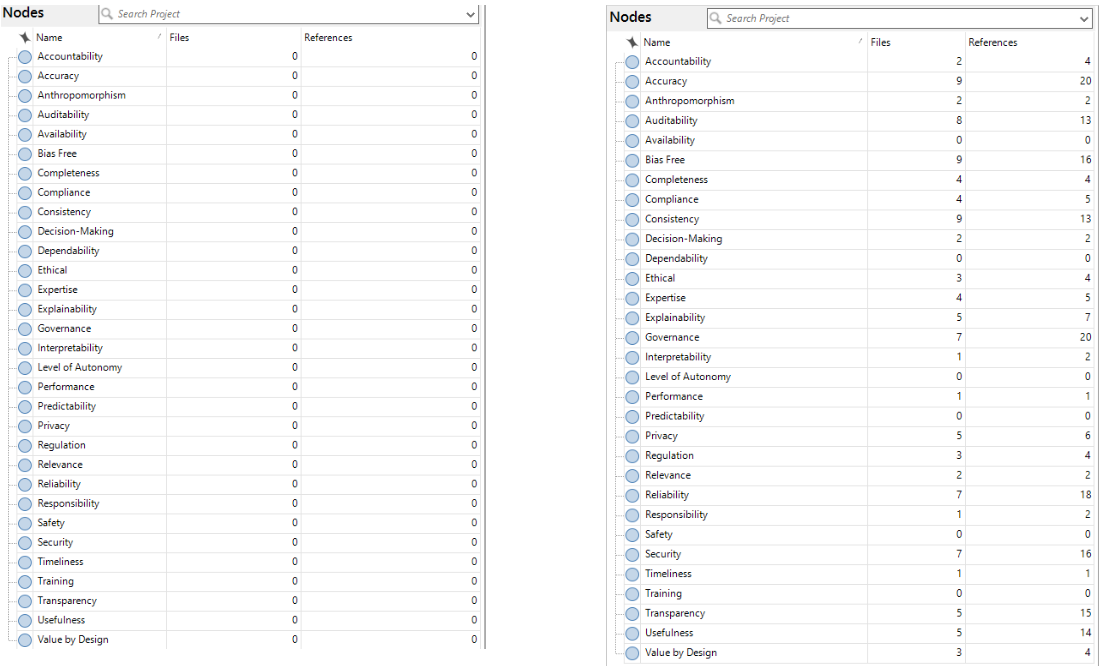


*From Risk assurance perspective, if a client wants to adopt technology like AI but they are not sure about the functionality of the technology and in that case they would be more interested towards "**Control by Design**" or like **Reliability by Design**. So at upfront, they want to know for sure that what's going into AI and what's coming out of AI is **complete, reliable, secured and accurate**.*

*There should be a need for people with **expertise** in order to understand the working of the tool, ability to understand or read the underlying code and the functionality behind that.*

Figure 31 - Snapshot showing the perspective of a manager towards trust in AI – Respondent 5

Now that, the summary of every interview transcripts were derived where some of the summaries had been presented as in figures 29, 30, 31 and the rest of them could be found in the Appendix A. To analyze the interview transcripts further (recorded interview conversation -> transcribed text), the method of ***content analysis*** will be used as it would enable the researcher to analyze a large amount of textual information, and systematically identify the important properties in the text, such as the presence of certain words, themes, concepts, characters (Sekaran & Bougie, 2016). In this research, the presence of trust factors or dimensions would be considered as the main properties in the text. To conduct content analysis, the text needs to be coded in terms of the factors of the model and dimensions of DQ which will be analyzed using conceptual analysis (establishing the existence and the frequency of factors and dimensions used in the text). The interview transcripts were coded using Nvivo qualitative software in order to draw important factors and conclusions from the transcripts. All the predefined factors of trust identified and studied from the literature were created as nodes and these factors will be marked as codes while carefully reading through the interview transcripts.



Name	Files	References
Accountability	0	0
Accuracy	0	0
Anthropomorphism	0	0
Auditability	0	0
Availability	0	0
Bias Free	0	0
Completeness	0	0
Compliance	0	0
Consistency	0	0
Decision-Making	0	0
Dependability	0	0
Ethical	0	0
Expertise	0	0
Explainability	0	0
Governance	0	0
Interpretability	0	0
Level of Autonomy	0	0
Performance	0	0
Predictability	0	0
Privacy	0	0
Regulation	0	0
Relevance	0	0
Reliability	0	0
Responsibility	0	0
Safety	0	0
Security	0	0
Timeliness	0	0
Training	0	0
Transparency	0	0
Usefulness	0	0
Value by Design	0	0

Name	Files	References
Accountability	2	4
Accuracy	9	20
Anthropomorphism	2	2
Auditability	8	13
Availability	0	0
Bias Free	9	16
Completeness	4	4
Compliance	4	5
Consistency	9	13
Decision-Making	2	2
Dependability	0	0
Ethical	3	4
Expertise	4	5
Explainability	5	7
Governance	7	20
Interpretability	1	2
Level of Autonomy	0	0
Performance	1	1
Predictability	0	0
Privacy	5	6
Regulation	3	4
Relevance	2	2
Reliability	7	18
Responsibility	1	2
Safety	0	0
Security	7	16
Timeliness	1	1
Training	0	0
Transparency	5	15
Usefulness	5	14
Value by Design	3	4

Figure 32 - Illustration of the codes created using Nvivo (Pre & Post coding)

Once all the transcripts were coded by mapping the factors to the nodes created, It was observed that factors like ***accuracy, auditability, bias-free, consistency, governance, explainability, privacy, reliability, security, transparency, and usefulness*** were identified to be the most important factors from the perspective of the actors interviewed. However, these factors have to be segregated towards the data and the model as there are many possibilities that these factors could mean important only to the data or the model or to both the data and the model. For example, factors like consistency, accuracy, reliability, usefulness, interpretability, etc. can be tied to both data and the model.

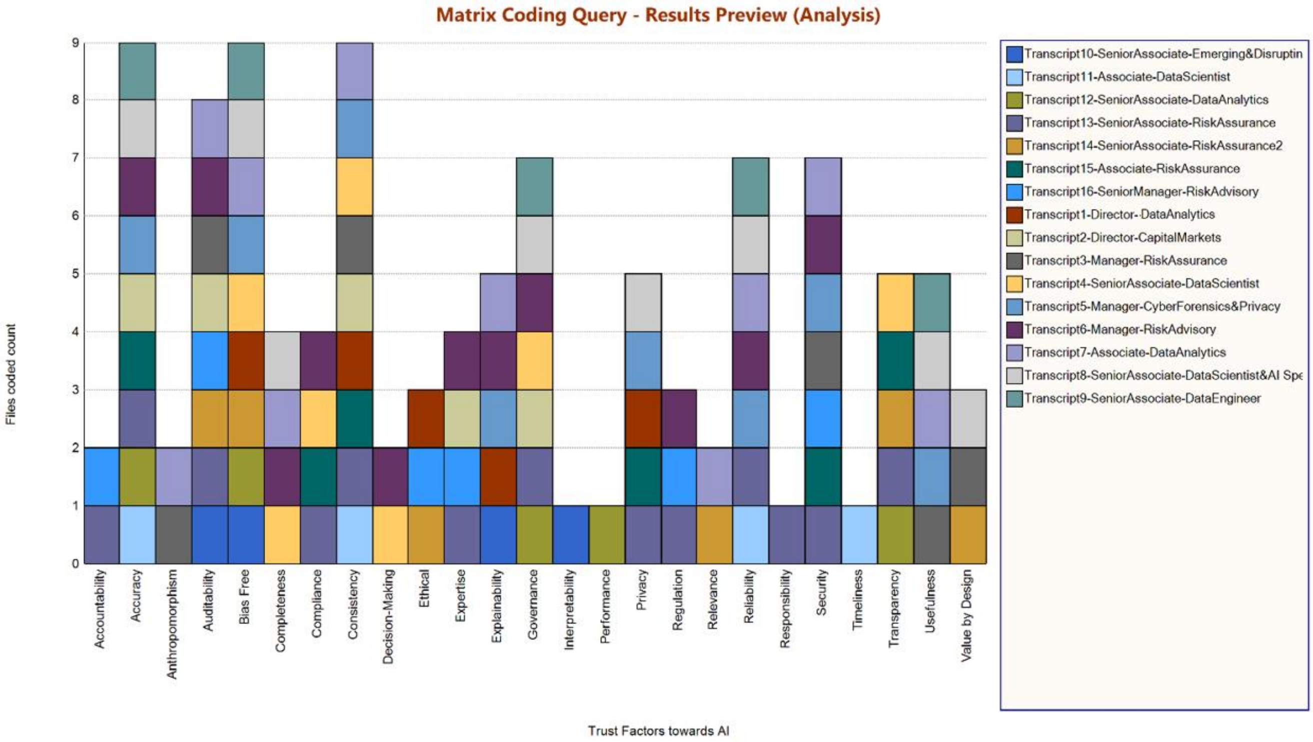


Figure 33 - Analysis of the trust factors using coding via NVivo

But, upon further analyzing the transcripts, many respondents highlighted that there would be a need for human trust factors as well which could also play a part in trusting AI. Factors like ***expertise, accountability, responsibility*** *and* ***even human bias*** could be an important trust factor towards the human involved in the development of AI. One of the senior associates from risk assurance department made a concise point on the importance of having expertise people during the development of AI,

If people are managing algorithm who are incompetent, then it's a risk from the perspective of risk assurance

(Or)

If people are working with data that gets fed into the AI is incompetent or haven't been trained on particular data set with respect to the specific context of the AI solution, then it's a risk again.

Now being clear that how factors have been perceived by the respondents from their perspective, they had to be clearly segregated not just towards the data, and the model but also towards the human actors. The below table shows the clear categorization of factors based on the actors perspective.

FACTORS	DATA	MODEL	HUMAN
Accountability		✓	✓
Accuracy	✓ (DQ)	✓	
Anthropomorphism		✓	
Auditability	✓	✓	
Bias-Free	✓ (DQ)	✓	✓
Completeness	✓ (DQ)	✓	
Compliance	✓	✓	
Consistency	✓ (DQ)	✓	
Ethical		✓	
Expertise			✓
Explainability		✓	
Governance	✓	✓	
Interpretability		✓	
Performance		✓	
Privacy	✓ (DQ)	✓	
Regulation		✓	
Relevancy	✓ (DQ)		
Reliability	✓ (DQ)	✓	
Responsibility			✓
Security	✓ (DQ)	✓	
Timeliness	✓ (DQ)		
Transparency	✓ (DQ)	✓	
Usefulness	✓ (DQ)	✓	
Value by Design		✓	

Table 11 - Categorization of factors based on the perspectives of the actors interviewed

One of the key findings from the analysis was that the respondents representing the data scientist, data engineer, data analytics consultant dint emphasize the factors like ***security, privacy or auditability with respect to data or model*** as an important factor from their stands. Though some of them agreed that security might at times be important depending upon the context of the client case, but they do it on a very low level and they assume or expect that it is taken care by other actors like managers and senior managers. Similarly, auditability was not seen as a critical factor and almost all the respondents from the data background stated that it would rather be important only after the implementation of AI and it would be more of the auditor's job to ensure that they are able to audit such system and provide a seal of trust to the stakeholders. Much to the contrary, respondents from risk advisory and assurance department tagged ***security, privacy, auditability, reliability, consistency, accuracy, explainability*** as some of the most critical factors towards the model. A manager from the risk assurance department gave a logical reflection on the security.

> *Tampering from outside in order to manipulate on how the AI behaves, how it decides and how it acts should be prevented at all times because when we find a useful purpose to actually use AI compared to a human or any other technology and if the security is not safeguarded or being breached, you lose the usefulness and mainly the trust in the technology.*

Most of these respondents also stressed some of the important dimensions in trusting data since data is the main driver or key to the model. Dimensions like ***reliability, accuracy, consistency*** were recognized as the main factors towards the data from the perspective of these actors (Risk advisory & assurance).

> *It is more about getting the data right and having an accurate process and when you find the cracks on the accuracy or maybe it's not reliable, you lose the trust towards the technology.*

It can be inferred that some of the factors were highly seen as important and in fact, many respondents from the data background emphasized the same factors and furthermore, respondents from risk advisory & assurance department stated the same set of factors towards the data. Figure 34 shows the list of trust dimensions towards the data that has been categorized in terms of priority based on the analysis from the interview. It has to be noted that factors like ***Governance*** and ***auditability*** may not contribute to the dimensions of DQ but those factors are believed to be important in order to place the trust around the data process and data life cycle stages. To clearly distinguish them, a conceptual map was derived as shown in figure 35 indicating the important dimension of DQ based on the respondent's perspective.

TRUST DIMENSIONS		**PRIORITY**	
Accuracy	🟩 **7**		
Timeliness			▨ **1**
Auditability			▨ **1**
Consistency		🟨 **4**	
Bias-Free	🟩 **5**		
Reliability		🟨 **4**	
Usefulness			▨ **1**
Explainability			▨ **2**
Completeness		🟨 **3**	
Governance		🟨 **4**	
Privacy & Security		🟨 **4**	
Relevancy			▨ **1**

🟩	Highly Important **(5-7)**
🟨	Moderately Important **(3-4)**
▨	Least Important **(1-2)**

Figure 34 - Categorization of trust factors towards data in terms of priority – Interview analysis

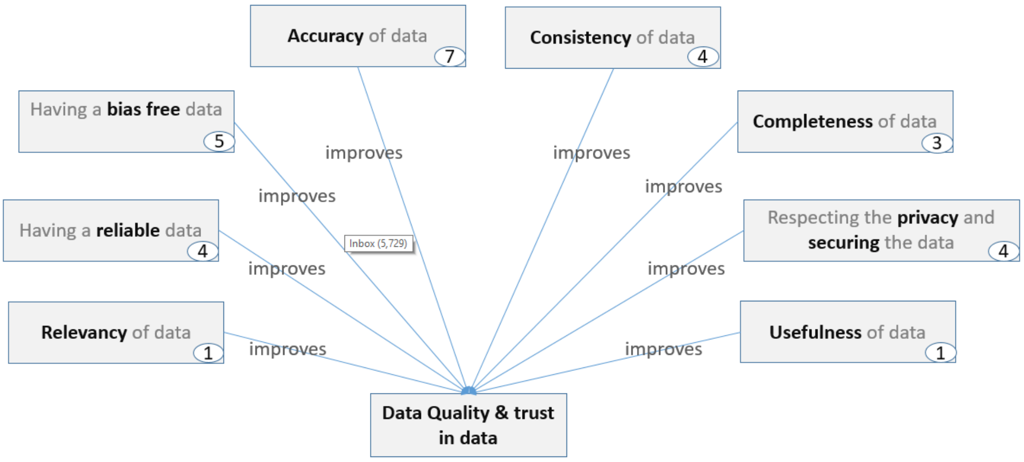


Figure 35 - Conceptual map showing the dimensions of DQ in the context of AI - Interview analysis

Similarly, with respect to the model, factors that were identified from the interview has been categorized with priorities based on the respondent's perceptions towards those factors and a conceptual map showing the factors influencing the trust in system and automation which is more of AI in this context is presented in figure 37.

TRUST DIMENSIONS	PRIORITY		
Auditability	7		
Bias-Free		4	
Completeness			1
Accuracy		3	
Consistency		4	
Ethical		3	
Explainability		3	
Governance		3	
Privacy		4	
Reliability		4	
Security	5		
Transparency	5		
Usefulness		3	
Value by design		3	
Performance			1
Accountability			2

🟩	Highly Important **(5-7)**
🟨	Moderately Important **(3-4)**
▨	Least Important **(1-2)**

Figure 36 - Categorization of trust factors toward AI model in terms of priority - Interview analysis

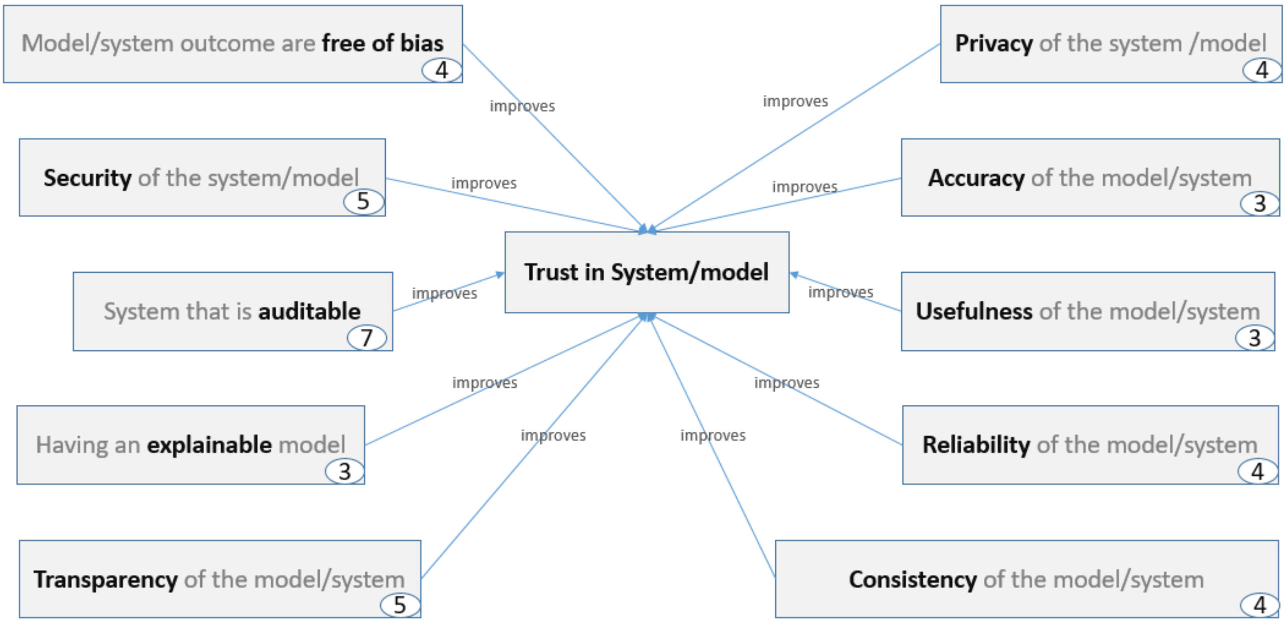

Model/system outcome are free of bias 4
Privacy of the system /model 4
improves
improves
Security of the system/model 5
improves
improves
Accuracy of the model/system 3
Trust in System/model
System that is auditable 7
improves
improves
Usefulness of the model/system 3
improves
improves
improves
improves
Having an explainable model 3
Reliability of the model/system 4
Transparency of the model/system 5
Consistency of the model/system 4

Figure 37 - Conceptual map showing the factors influencing the trust in AI model – Interview analysis

From the above figure, it can be derived that there has been increasing attention over the factors like governance, bias-free, consistency, reliability, etc. towards the data and model as a whole. It is also understood that how these actors value their importance on having data that is free of bias and establishing governance not just around the data but also over the model. The findings also visualize a few trade-off factors with respect to the data and the model. (1) One of the factors, ***explainability*** was considered to be least important as a trust factor towards the data while on the other side, respondents expect the model to be explainable whenever the model throws out a recommendation or decision and in such cases, explainability would be really important and it was even hinted that if such models need to be explainable, it needs to be transparent as well, as that could help the auditor to examine such system critically. So, factors like explainability and transparency are interrelated and were highly regarded as crucial trust factors of the model. (2) Secondly, ***Auditability*** was not seen as a crucial factor with respect to the data but from the model perspective, it was given a high priority especially from the respondents of risk advisory and assurance. It was believed that auditability is one of the important aspects of trusting AI as it involves a thorough examination of the system & its associated boundaries. People from risk advisory & assurance department even cited that those working on data have to consider the importance of audits in mind and ensure that every change made with respect to the data are documented. (3) Finally, ***security & privacy*** was equally considered important in data & model, but it had high priority mainly towards the model than the data. It was emphasized by the managers, consultants from risk advisory, assurance, and security practice. it is believed that for any digital technologies like IoT, Big data, AI that uses data as a core input, security & privacy is something that needs to be validated upfront and when AI is going to be fed with large volumes of data or when it is going to utilize the user information as input, it requires critical assessment and establishing robust security and privacy frameworks not just around the model but throughout the entire framework which includes data as well.

To summarize the main findings based on the analysis of the actors perspective in trusting AI, it can be concluded that,

- Accuracy, consistency, completeness, security & privacy, usefulness, relevancy, reliability, objectivity (bias-free), etc. were identified to be the dimensions that constitute the DQ. Among these factors, accuracy, consistency, bias-free data, reliability, completeness was considered to be the crucial factors in improving the DQ and thereby improving the trust in data based on the perspective of the actors.

- To be more specific, actors like data scientist, data engineers, data analytics consultants recognized consistency, accuracy, completeness, bias-free, reliability, etc. as the important dimensions of DQ and governance as the most critical factor in order to trust entire data process and its lifecycle. If data can be trusted by assessing the above dimensions, the outcome of the model will also be trustworthy. Though these actors didn't underline much of the trust factors towards the model, it was understood that if the model that has accurate, reliable, consistent, bias-free data and is being trained with such data, the outcomes of the model would also be reliable, consistent, and bias-free. Data really makes a difference and it can directly influence the trust of the technology.

- People from risk advisory and assurance regarded auditability, security, privacy, reliability, bias-free, transparency, explainability, governance, ethical, and usefulness as the important factors in trusting the model. It was also believed that most of these respondents would be involved either in the initial phase (advisory) or at the final stage (assurance) during the development of AI and they would mainly see the above factors as vital in trusting the resulting AI model.

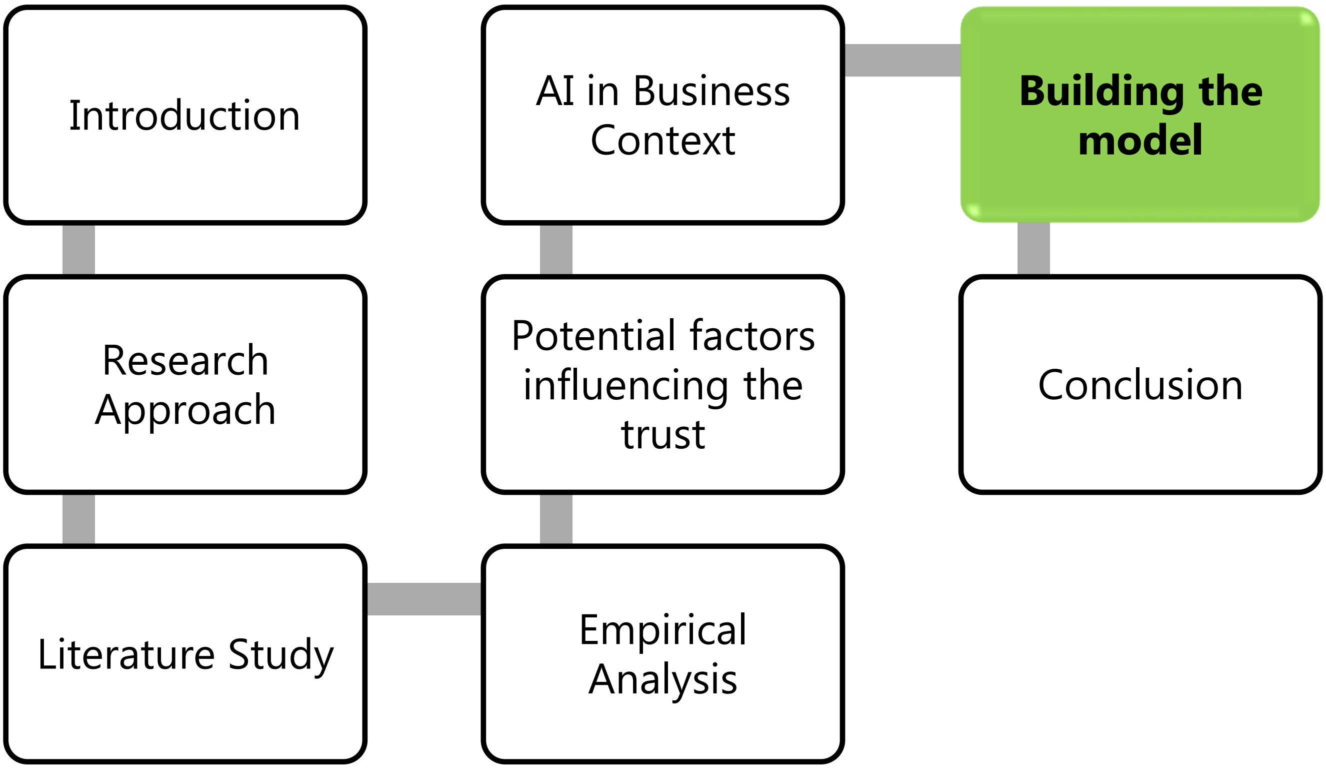
Additionally, the findings gave **unanticipated** insights and perceptions,

- It was realized from the interviews that the role of human trust factors has to be taken into account as that can also influence and improve the trust in AI. Factors like **expertise, accountability, responsibility, and human biases** were some of the crucial identified factors based on the interviewee responses. For instance, the expertise of the developer over the technology would be crucial in AI development. If the developer lacks expertise in choosing the right model or not being trained on handling the data, the resulting AI solution would be futile and unreliable. This not only applies just to the developers but also other associated actors like auditors who should be proficient enough to audit the AI systems before providing a seal of trust to the clients. Second, establishing clear accountability and responsibility standards with respect to human actors involved in the AI development would influence the trust in resulting AI technology. Though these factors are undoubtedly crucial trust factors of human actors, it has to be admitted that this research won't elaborate them in detail as these factors would require further study and analysis and this was recognized only when the actors were interviewed. Identifying the trust factors of human actors was not the main purpose of this research as specified in the scope of this research in chapter 1.7 (human characteristics). More specifically, the focus of this research is to find the essential trust factors of the data and the resulting AI model only.

- The analysis from the interview also reflected on the importance of creating value by design upfront, inclusiveness as a major trust factor in the early design phase. This means involving and engaging the right people upfront during the initial phase of the design, whether the targeted group of the application has been involved in the design phase, etc..

7

BUILDING THE MODEL



Using the literature and the findings generated from the interview with relevant actors as the main source, an initial version of a trusted AI model would be developed. However, the model would initially be drafted with detailed phases involved in the development of AI and what kind of trust factors would be crucial to be considered at each phase would be identified. Since there is no existing or concrete trustworthy AI model available in the academic research that has explicitly defined the main trust factors towards AI, a relatively new model has to be developed upfront clearly segregating the factors towards the data and the model. This drafted model created would then be evaluated with some of the relevant experts in AI and make sure that their feedbacks and reflections are taken into account before realizing the final version of the trusted AI model.

Additionally, the final proposed model would be compared with some of the themes laid as trust factors by the research institutions & EU Commission to ensure that the model has considered those themes as well. They have been considered as the additional milestones for this research.

As the EU commission has recently published a document on ***ethical guidelines for trustworthy AI*** written by the high-level expert group on AI (European Commission, 2019), It will also be made sure that the resulting model is complying with the factors presented in the document as that could add high value to the model.

Many AI research institutions like ***Partnership on AI, Future of Life Institute, AI NOW Institute, AI for Good, Algorithmic justice league***, etc. has been emerged to mitigate the possible risks & solve the pressing issues circled around AI. These institutions major themes include fighting biases, addressing the social implications of AI, promoting diversity and inclusions, making AI fair, explainable and accountable, etc. It needs to be ensured that the proposed model in this research has considered and addressed most of their themes that could possibly contribute new knowledge and insights to those institutions as well.

It has to be noted that there are no specific standard stages defined in scientific research with respect to AI development. Perhaps it would change depending upon the context of every organization's workflow. This research has used Guo's (2017) study where he identifies several main stages involved in the development of machine learning (one of the dimensions of AI) and his study was used as the main source of reference in identifying the probable phases involved in the development of AI with regards to this research along with the interview findings from the experts (data scientists, directors, AI experts). Most of the phases like Data acquisition, Data preparation & validation, Model selection, Model Training & Testing, Model Validation & deployment were in fact retrieved from the study of (Guo, 2017). Further, it was also realized that these identified stages were more or less in accordance with many leading tech companies process flow in developing AI-based products or solutions. Next to the identification of phases are the trust factors which is the prime aspect of this whole research. Although various literature was analyzed extensively in order to determine the potential trust factors of the data and the AI model at the first instance, some sources of literature were recognized to provide a major input to this research. For instance, most of the DQ dimensions were retrieved from the studies done by (Bovee et al., 2003; Janssen et al., 2017; Lee et al., 2002; Pipino et al., 2002; Wand & Wang, 1996; Wang & Strong, 1996). Accuracy, completeness, relevancy, consistency, reliability, interpretability, etc. were some of the common dimensions identified from these sources. In determining the potential trust factors of resulting AI model, not many academic journals were found in the context of AI. Instead, various white papers published by tech companies and themes emphasized by AI-based research institutions were used as one of the main references and mainly the literature that was done on the factors of trust & trust models in automation, human-robot interaction, etc. were used. The below figure 38 clearly depicts the sequential flow of methods as inputs required for building the trusted AI model. The figure also presents some of the main sources that were used in determining the trust factors of the data & model. More specifically, the method of arriving at the model is showcased in figure 38.

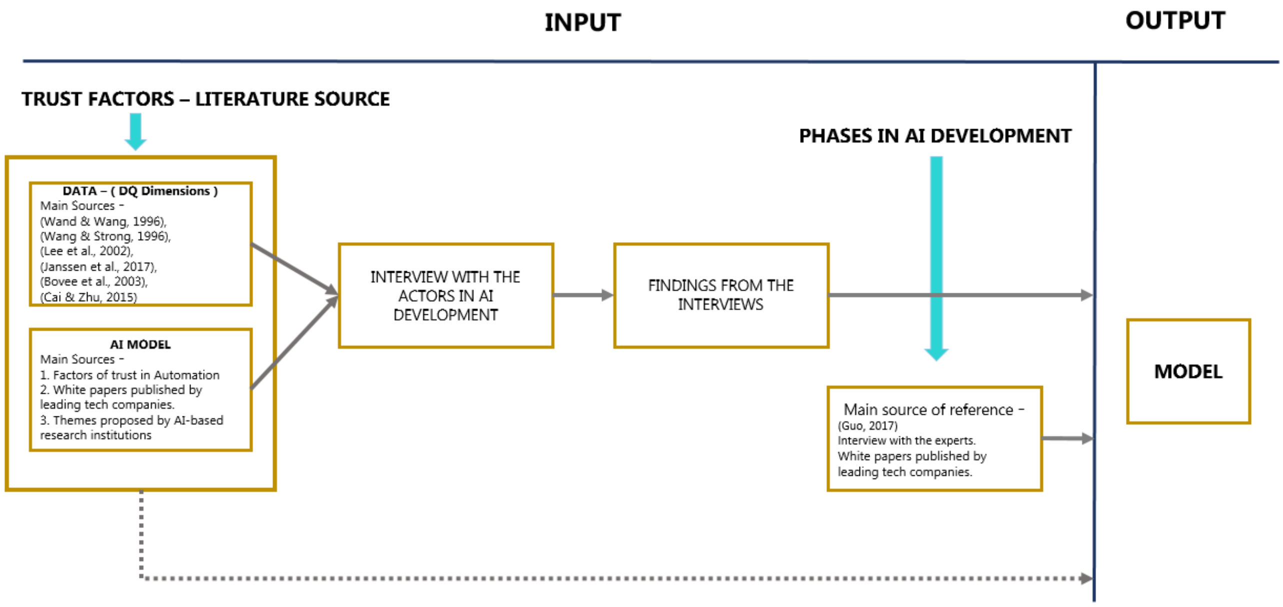

INPUT
OUTPUT
TRUST FACTORS – LITERATURE SOURCE
DATA – (DQ Dimensions)
Main Sources -
(Wand & Wang, 1996),
(Wang & Strong, 1996),
(Lee et al., 2002),
(Janssen et al., 2017),
(Bovee et al., 2003),
(Cai & Zhu, 2015)
AI MODEL
Main Sources -
1. Factors of trust in Automation
2. White papers published by leading tech companies.
3. Themes proposed by AI-based research institutions
INTERVIEW WITH THE ACTORS IN AI DEVELOPMENT
FINDINGS FROM THE INTERVIEWS
PHASES IN AI DEVELOPMENT
Main source of reference -
(Guo, 2017)
Interview with the experts.
White papers published by leading tech companies.
MODEL

Figure 38 – Inputs for the trusted AI model

7.1 MAIN PHASES

The main phases that are identified will cover right from problem identification to the monitoring of the AI model or systems post-deployment. The phases included are: (1) Problem/Improvement Exploration, (2) Human-Centered Design, (3) Data Acquisition, (4) Data Preparation & Validation, (5) Feature Selection, (6) Model Selection, (7) Model Training & Testing, (8) Model Validation & Deployment, and (9) Model Monitoring. Each of these phases would be given an overview and the emphasis would be mainly placed on the trust factors which are likely to influence in each phase. The purpose of introducing these phases is to provide a complete background of how AI/ML would work in reality and what would be the trust factors that need to be validated in each of these phases in order to build and improve trust in AI.

It is important to recognize that phases like problem/improvement exploration, human-centered design, etc. won't be supplemented with enough details. The purpose of mentioning these phases is to make the model complete & provide a clear view to the readers in realizing the actual phases involved in the development of AI.

Problem / Improvement Exploration - It is assumed in this context that the relevant stakeholders (Client and technology creators) have already established a good relationship of trust & won't have any issues in identifying the known problem or possible improvements. It was understood from one of the interviews conducted with an AI expert that in most of the times, PwC goes to the client to understand the challenges that they are dealing with, translate them to the technical requirements and determine which technology can provide value to the clients.

Human-Centered Design - Over recent times, academic researchers, AI research institutions and industries are trying to address the issues of AI especially in the design stage of AI. Lack of diversity in AI workforce, not engaging the right stakeholders in the design phase, and not involving the right human in the AI loop & workforce has been a major problem for most of the risks evolved around AI. It has already been reported that there has been a huge diversity crisis in the field of AI across gender and race (Crawford et al., 2019). The example of AI-based recruiting tool mentioned in chapter 1.2 is a clear cut example on the importance of considering the gender and race as a high priority especially when such application is deployed in recruitments. It won't alone be the case of men over women or white over blacks but also the trans people who would be affected as well, as there is no public data on trans workers or other gender minorities (Crawford et al., 2019).

The research would place touchpoints in the phase of human-centered design by identifying some of the important factors that need to be considered during the design phase. It was even emphasized by several respondents during the interview, that involving the right people upfront during the design will largely influence the trust in AI as the majority of the risk encircled around AI can be mitigated to a greater extent. The table represents the list of factors that should be reflected in the design phase.

FACTORS	INDICATORS
Inclusiveness	• The potential AI users, developers, representative of industry sectors who might be impacted by AI are included. • Encouraging the involvement of key actors like data scientist, risk advisors, policy regulators to understand the challenges, requirements in the early design phase. • Domain-specific functional specialists are included apart from the techies to make sure that the resulting model is used effectively in the specific domains.
Diversity	• Defining the target group and ensuring whether all the representative of those targeted group is involved during the design. • A diverse set of representatives are present during the design phase in order to encourage the diversity of opinions.
Value by design	• The proposed solution is actually giving value to the clients. • The solution proposed would be the best solution to the business needs and to the society (For example, AI chatbot).
Societal & Environmental well being	• Seeking guidance from the experts, the target group of the solution in order to solve the problem and promote "AI for good". • Emphasizing the importance of privacy, security, and safety with the relevant actors in the early design phase. • Collaboration with private and public sector institutions when it comes to AI's Societal important . • The proposed solution is made to ensure that it would benefit the human and not harm them in any instance.

Table 12 - Trust factors in Human-centered design phase

Data Acquisition - Data Acquisition is the prime phase to establish initial trust. This is the phase where the developers of data such as data scientist, data engineer & stewards get involved and start gathering the data. There are several ways to acquire the data & it depends upon the context of the client case. There would be instances where the data can be obtained directly from the client (or) possibility of getting the data from an external party or data provider (or) the public data sources. There would be scenarios where one needs to consider the client data as well as the third parties data. An example could be, Forecasting the customer purchase as that could help the client to plan their logistics, supply chain and delivery and if the application has to consider the weather conditions in order to plan their delivery, the developer would need data from weather-based application providers as well (third party), as that could help to improve the overall accuracy of forecasting.

The importance of trust towards the data initiates from here where the dimensions of DQ would serve as a base in assessing the quality of data and placing the initial trust over such data. Possible dimensions of DQ that needs to be assessed at this phase are the reliability of the data, accessibility of the data, availability of the data, completeness of the data, interpretability of the data, consistency of the data, and having a bias-free data. These dimensions have to be assessed based on the below principles or indicators.

DIMENSIONS	INDICATORS
Reliability	• The contents present in the data set is credible enough to process. • The data has been generated from a trusted source.
Accessibility	• The data is readily accessible or retrievable from the intended source. • A higher degree of openness in accessing the intended data.
Availability	• Data required is readily or physically available from the target source. • The main features that are relevant to the solution are available in the data.
Completeness	• The mandatory & main attributes in the data are not null. • The data has all the possible states relevant to the user population. • The data is completely representative of the real-world state.
Interpretability	• Data is conveying intelligible and meaningful information in the first instance. • Data obtained has appropriate language, symbol, and units. • The language & metrics used in the data is known to all actors working in the data.
Consistency	• The source data is consistent even after a certain point in time. • The physical instance of the data is in accordance with some standard format. • The value and entries in the data are the same in all the case.
Bias-Free data	• Data received has more sample towards a particular category or label which requires validation. • Identifying a collection or selection based biased data in the first instance.

Table 13 - Dimensions of DQ in Data acquisition phase

Data Preparation and Validation – Once the initial trust is established through Data acquisition phase, continuous trust has to be maintained in the phase of data preparation and validation, the phase which involves a series of steps in processing the data. This is also the phase where the processed data can be split into two parts where the first part is the training data that would be used for training the model and the second part is the testing data that would be used to evaluate the performance of the model once the model is being trained (Guo, 2017). In either case, the dimension of DQ has to be assessed in this phase.

The data processing involves four main steps: Data Cleansing – removing the bad data, duplicates, dealing with missing values, and data type conversions, Data Filtering & Wrangling - Discarding the irrelevant data and mapping the data to a format with the intent of making it more appropriate and valuable, Data Labeling – Label the data based on the categories if present, Data Visualization – to identify the trends, relationships between the variables and perform analysis. Once the processed data is ready, it has to be divided as testing and training data. Now, the dimensions of DQ has to be seen before selecting and training the model. The probable dimensions of DQ in this phase would be:

DIMENSIONS	INDICATORS
Relevancy	• The processed data is relevant to the specific domain and purpose of interest in a given context. • Processed data contains the required variable in the right form and a representative of population interest.
Usefulness	• The processed data is in line with the context of the goal. • A useful value is identified in the processed data that can improve the overall usefulness.
Accuracy	• The processed data hasn't lost the structure during the data processing stages. • The processed data is certified and free of error. • The processed data is accurate, objective & is corresponding to a real-world context.
Bias-Free	• The biased data identified during the data acquisition and processing phase has been removed. • Data contains all the possible representation of the subject, object and has no inclinations towards a specific object or a thing, i.e no traces of biases, prejudiced, and partial data were found.

	• Determining the cause if any variance is detected so as to avoid biases.
Consistency	• The processed data and actual data source are consistent, have no contradiction and are compatible with the previous data. • The attribute values in the data have no ambiguities after data processing.
Privacy & Security	• Identify and protect the core strategic data assets. • Access to final processed data is restricted & secured from unauthorized access. • Processed data has been encrypted & anonymized if any personal or confidential information is present. • Processed data is used fairly and respects the privacy of the user's data. • The processed data is in compliance with GDPR and other data protection policies.

Table 14 - Dimensions of DQ (processed data) in Data preparation and validation phase

Feature Selection– At this point, it has to be ensured whether the best out of data has been derived as that can help the model to produce the best results. So, selecting the right features from the data is of paramount importance as it reduces overfitting of the model, improves the accuracy of the results by discarding the data that is misleading, and reduces the training time by having only important data points that make the algorithm train faster (Shaikh, 2018). The process of selecting a subset of relevant feature from a large pool of features present in the data is referred to as feature selection. The possible dimensions would be completeness, accuracy, and relevancy of the features.

FACTORS	INDICATORS
Completeness	The chosen features are representative of the entire population.
Accuracy	The chosen feature can help the model to make correct predictions, decisions, and recommendations.
Relevancy	The chosen features are applicable and addressing the actual context of the goal.

Table 15 - Trust factors in the Feature selection phase

Model Selection – Once the trust is established towards the data by mainly analyzing the dimensions of DQ as trust factors in the last three phases where data has been the core driver, a relevant model has to be selected. It has to be ensured that the chosen model would work well with the data which is about to be fed into the model. The main factors that can influence the trust in this phase would be accuracy, usefulness, bias-free, explainability, interpretability, and security.

FACTORS	INDICATORS
Accuracy	• The chosen model can produce correct predictions, decisions, and recommendations results if trained properly.
Usefulness	• The chosen model would be in line with the actual context of the goal.
Bias-Free	• The chosen model is not pre-trained to show any biases and can work well if trained properly.
Explainability	• The model can provide some explanation for every certain output.
Interpretability	• The result that a model can produce is interpretable with respect to the data.
Security	• The chosen model can be resistant to malicious training.

Table 16 - Trust factors in the model selection phase

Model Training and Testing– Having selected the relevant model based on the data analysis, the model would require training and it would be trained based on the training data that was kept aside during the earlier phase of data preparation and validation. The training data are used in small increments to improve the ability of the model to predict/recommend/do actions correctly. Once the model has been trained, it has to be tested. Test data at this stage would come into the picture and these test data should be different from the training data. If the training and test data are the same, the model will easily predict the results as it had the answers in advance.

The purpose of using the test data is to evaluate the model against the data that has never been used for the training (Guo, 2017).

FACTORS	INDICATORS
Transparency	• The training and testing data is open to be reviewed by domain experts.
Bias-Free	• The potential skews found during the training has been addressed. • Any skews observed during the testing were identified and addressed. • The training and testing data are free of biases.
Security	• The model is making fair predictions with the data being trained. • The model is neither using the test data maliciously nor being fed with malicious data and is making fair predictions with the test data.
Usefulness	• Tuning has been made to improve the overall usefulness of the model.
Accuracy	• Testing data & training data has samples that represent all the targeted subjects, things, objects, etc. • The model is able to produce desired results based on the training. • The model is able to produce the same set of desired results when tested with new data (test data).

Table 17 - Trust factors in the Model training and testing phase

Model Validation and Deployment- The model that has been trained and tested has to be validated and deployed at this time. The main push for validation is to safeguard the model and to ensure that the model has taken into the considerations of some important themes that would serve as trust factors. Factors like explainability, AI model being ethical, and the usefulness of the model, establishing the accountability standards, completeness & reliability of the model, privacy & security of the model would be some of the important factors that need to be taken into account at this phase.

FACTORS	INDICATORS
Explainability	• The AI model is able to explain every decision, recommendation, the prediction made. • The decision made by the system can be understood and traced by human beings.
Completeness	• The model is representative of the entire population.
Ethical	• The AI models or system are complying with the ethical norms and social values which also includes the human behavior involved in designing, developing and implementing, and AI as a virtual person.
Reliability	• The AI models are performing as intended. • The AI models are working properly when a new range of inputs is being applied.
Bias-Free	• The outcomes from the AI model is not inclined toward a particular subject. • The resultant AI model is being fair to all the targeted population and doesn't encourage unfair biases.
Usefulness	• The resulting AI model had addressed the challenges and is providing value to the clients, society and the stakeholders involved.
Privacy & Security	• The data used by AI systems or model is not used unlawfully or unfairly against the users. • The AI systems are safe and secured and are not vulnerable to tampering or compromising the data they are trained on. • AI models are protected from being exposed to unexpected situations. • The AI models developed are limited to the context of the goal. • The AI model respects the user privacy of information.

Table 18 - Trust factors in model validation and deployment phase

Model Monitoring- The final stage of the phase is the monitoring phase and in this phase, the model has to be closely monitored especially when the models are deployed in critical environments. One of the keys to monitor the model is to understand the potential weakness as it would give a better reality of what the models are actually doing, how are they considering the inputs and where are they failing, etc. when it is being deployed in real-time. This would be easy for the data-scientist, developers to improve the model accordingly by tuning and making adjustments to the model.

FACTORS	INDICATORS
Performance	• Identifying and addressing the potential weakness can improve the performance of the model. • Tuning and adjustments are made to improve the performance of the model.

Table 19 - Trust factor in the model monitoring phase

The model entitled with 9 main phases has to consider the above possible factors and dimensions to establish the trust over the data and the resulting AI model. However, the model is certainly not complete as it needs to be englobed with few important themes serving as trust factors and these factors largely influence ***overall*** trust in AI. They may not directly contribute to the data or the model but they are critical to be in place in order to place trust around the process of data and the model. Auditability and Governance are those factors which were regarded as highly important based on the findings from interviews and increasing attention that was paid from research academia and institutions.

FACTORS	INDICATORS
Auditability	• Every change made with respect to datasets that have been documented is verified. • A process on how data was received, analyzed, cleansed, filtered, and labeled are documented and verified. • Assumptions made on the data to improve the usefulness of the data has been documented and reviewed. • Possible risks and opportunities that were encountered and how they were addressed is documented. • The results produced by AI systems are compared with the actual results. • Reviewing the output of the AI model and the meaning derived from those outputs. • Verify by interviewing the ones responsible for developing AI models that they are able to understand and explain the AI black box data. • Assessing the already existed data from the client or other third-party vendors to validate the usefulness of the data for requirements
Governance	• A clear line of accountability, responsibility standards is clearly established. • Know the responsibility of the data & model owners at each and every phase involved in the creation of AI. • Ensuring that human is involved in every phase of the model as well as to oversee the overall activity of the AI systems. • Data protocols outlining 1. Who can access the data and under which circumstances. 2. How the data are being handled and how is it being protected. 3. Whether the data is in compliance with GDPR & data protection policies. 4. What kind of biases were identified, how were they solved? 5. Were there any measures been taken to inform the relevant stakeholders when a major biases were notified? etc. should be put in place.

Table 20 - Trust factors around the trusted AI model

The initial version of the model, therefore, contains nine phases with the possible trust factors that need to be assessed were identified in each phase and the entire model is englobed with two core trust factors that are necessary to be in place as that can help to improve the overall trust in AI. The initial version of the model can be found in Appendix E. Since the model is comprehensive, it needs an evaluation from the experts and the relevant actors who are very likely to be involved in the development of AI. The evaluation of the model would certainly help to improve the utility of the model and therefore, a final version trusted AI model can be derived.

7.2 MODEL EVALUATION

The main purpose of the evaluation is to realize the value of the model by seeking suggestions and feedback from the experts who are involved in the specific application domain. Evaluation with such experts would help to answer the question of, ***how well does it work in reality?***. Normally, different criteria are used during the evaluation. For instance, in evaluating a concept, it would tend to involve completeness, simplicity, elegance, and understandability while the model is evaluated in terms of fidelity with the real world phenomena, completeness, level of detail, robustness, and internal consistency (March & Smith, 1995). This research would tend to follow the strategy in terms of the model only. Two qualitative based approaches would be used to evaluate the model: evaluating the model by conducting an interview with experts and relevant actors involved in the development of AI (Data Scientist, Manager – Risk Advisory, Risk Assurance) and evaluating the model with a case study.

EVALUATION OF THE MODEL USING THE INTERVIEWS

To evaluate the level of detail in the model & the level of usefulness in the real world context, interviews were conducted and by going with the belief that respondents are already aware of the research being carried out as some of the respondents who were involved in the initial research were interviewed back again as a part of the model evaluation. However, the respondents selected in this case were made to ensure that they are either quite aware or involved in most of the phases that were identified during the model building. So, the targeted people were a data scientist, consultants & managers from the risk assurance and advisory department. The interviews were constructed as follows, giving them a short background of the study again to the respondents, explaining the initial version of the trusted AI model, reflections and recommendations over the model by the experts.

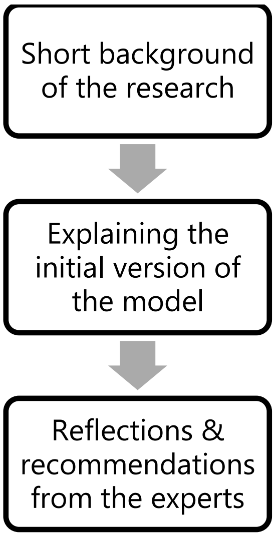

Short background of the research
Explaining the initial version of the model
Reflections & recommendations from the experts

Figure 39 - Evaluation method – Interview strategy

ACTOR 1- DATA SCIENTIST- To start with, the first respondent interviewed was a data scientist who has a wealth of experience in applied statistics, data mining, data analysis, machine learning, Hadoop, etc. and is currently engaged in working on warehousing and & analyzing large data sets.

The data scientist, after analyzing each phase of the model that was presented, gave a lot of interesting insights and reflections over the model. The data scientist was advised to give reflection over every phase of the model but realizing his expertise over the data, he started giving his feedbacks from the third phase of the model which is the data acquisition phase.

Data acquisition– The data scientist acknowledged that factors like ***completeness, reliability, and consistency*** of the data would be the most important factors regardless of the context of the case. He would see them as the main factors to trust the data in the first instance. He even stated that finding biases, at first sight, would be very challenging as most of the times, the data received would be raw or unformatted & unstructured data. Secondly, accessibility would be a factor to be considered but then he cited that they haven't faced any issues in accessing the required data as most of the time, they go to the client's place, say the requirements and the clients provide the data. However, the data scientist made an insightful comment,

Accessing the data is not a major challenge in the above context. But, when we need to work on AI-based solution where we would need some public data sources, then there would be a major hurdle in accessing the data as some of the data sources that we are looking for, will be blocked or we would be able to access some part of the data contents only. In such a case, we care about ***accessibility****. But I think in general it's still a factor that needs to be considered but I would rather first put my attention on* ***completeness, consistency, and reliability of data*** *for making predictions. If the data doesn't even represent the labels of the final outcomes, then you can't really make predictions out of it.*

Data preparation and validation - is the phase where the data scientist cares about the **biases** in the data, once the data is cleaned and processed. ***Consistency*** of the data as they check whether the data is still representative after cleaning them, and finally the ***relevancy*** of the data. As factors like privacy & security were identified in this phase, the data scientist made an interesting remark,

Even factors like privacy and security are important in this phase, but the managers or directors usually make this call. Normally when we use to build the model, we try to make sure that we just came up with ideas without any hindrances and then if the managers hint that there would be a possibility of a security or privacy issue, they would then kind of let us know on that and if they don't say anything about that, we will not care about those factors

Additionally, the data scientist recommended that factors like **privacy, security or compliance** are something that can be tied to the governance as such factors play a big role inside the governance.

Feature selection – In this phase, the data scientist valued relevancy and completeness as an important factor than accuracy.

I don't think I care about accuracy at this point at all because we mainly care about accuracy in the ***model*** *as it is very too hard to know in this phase if the accuracy of the model will improve or not, based on the features selected. Rather, we would see* ***completeness*** *for sure and it is super important. When you start the feature selection and you select five to ten feature as important from the data but they represent 60% of the data, the model would predict well only for the 60 %.*

Model Selection – Factors like ***accuracy*** and ***bias-free*** were recognized as important and additionally, the data scientist made an interesting remark over the importance of ***interpretability*** at this phase,

Clients actually care a lot in having a model that is ***interpretable*** *but we as data scientist won't value high attention to interpretability as we see at the end whether the model can produce the best results from the data. But, since we see a lot of news on the outside world saying the AI model should be interpretable, we would need to build such model that can be interpretable but at the same time produce efficient results.*

Model training and testing – ***Bias-free, usefulness, and accuracy*** were recognized as crucial factors by the data scientist during the training and testing of the model.

Model validation and deployment – Same like ***Interpretability*** that was emphasized by the data scientist in the model selection phase, having a model that is ***explainable*** is of paramount importance. If the model that is selected for training are interpretable, it gives more space for the model to be explainable and transparent. He acknowledged by saying that,

> ***Explainability*** *is one of the factors that we as a data scientist wouldn't have cared much about but, now most of the client wants to know the explanation behind the models result. In such stances, we have to see explainability as a crucial factor.*

Reliability and ***usefulness*** of the model would be important at this phase in order to make sure that the model is providing the actual value and solution to the clients that they expected. Considering bias-free at this point would be least important as,

> *We believe that we have removed all the underlying biases that were there and you have the model which is as bias-free that we can expect from the final model.*

Model monitoring - Performance would be the only factor that was present in this phase which was agreed by the data scientist as well. They would closely monitor the performance of the model and see if the model would require any tuning or adjustments to improve the accuracy and performance of the model.

The data scientist also suggested that governance with respect to the model can have factors like accountability, responsibility, regulation, ethical, privacy & security, etc. and these factors can be tied to the governance separately rather than linking to the phase as governance and auditability serve as main trust factor around the entire phases.

Finally, a follow-up question was asked regarding the perception of data scientist towards the initial version of the model, the data scientist responded that this model is quite detailed but not complex and it really makes sense to the current context of AI where trust is always a big question. If such kind of model can be shown to the clients upfront, they would really be interested to speak further about the possibilities and maybe the model can accept modifications based on their requirements.

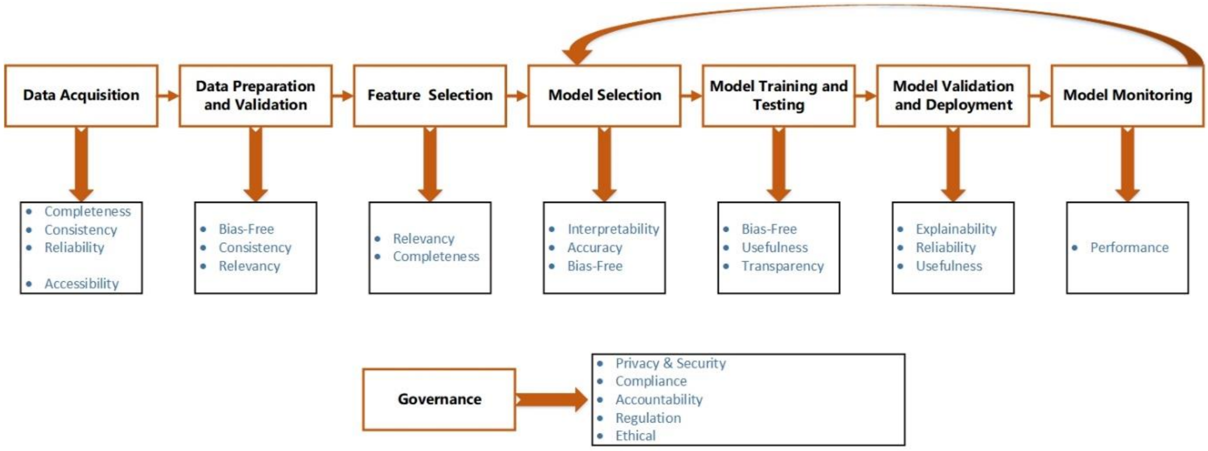


Figure 40 - Evaluation of the model – Insights and recommendation by a data scientist

To further realize the potential value and the usefulness of the model, Interviews were further conducted with the managers from the risk assurance and advisory practices. It is believed that these managers play a significant role in the governance and audits which are considered to be important factors in this research.

ACTOR 2- MANAGER (RISK ASSURANCE)- One of the managers from the risk assurance was more comfortable in providing his insights and suggestions towards the beginning and the last part of the phase as they are more engaged with the clients and other stakeholders at the first occurrence and they would be the people who would critically assess the solutions just before the deploying of solutions in the production or in the client space.

Human-centered design – In this phase, the manager made a concise remark that factors like ***diversity, non–discrimination*** are all part of ***inclusiveness*** and can be attributed to inclusiveness as one factor which is definitely an important factor as it can help to improve the adoption of AI. Secondly, value by design would be a crucial factor in this phase.

Model validation and deployment - Factors like the ***reliability*** of the model would be most important from the perspective of the manager as at the end, it needs to be tied to the value by design to ensure that the model is actually doing based on what was being designed initially. The manager also reflected that **reliability** and ***usefulness*** are twofold and so in that context, usefulness is also a prime factor

Reliability and usefulness would be the same because, without the reliability of the model, it wouldn't be useful.

Besides that, AI should make ***ethical*** decisions as it can certainly help to improve its adoption not just by the businesses but also the society as a whole.

If AI is in compliance with regulations but it makes unethical decisions, then the model will not be a great success nor the technology as a whole.

The third most important factor would be the compliance and the manager made a clear remark that,

It might be less weird to say compliance as a third factor especially from the perspective of risk assurance but without the other three factors reliability, usefulness and ethical, compliance would not be even present for discussion and it would be meaningless.

The manager also endorsed that auditability should be present in the model validation and deployment phase because

If the auditability is done only after validation, you might not know whether the model is auditable or not and ideally your monitoring won't be a success. However, it should also be a part of model monitoring.

As a key & final reflection, the manager realized plenty of positives from the model as it has two different sections, like one is driving towards the value of the model by seeing factors like completeness, reliability, usefulness and the other ones are geared towards more control mechanisms like is the model actually controllable, can we actually make sure that the model is doing everything within rules that have been set, within the right security parameters. So, it is a mix of value creation and protection of the model.

The model is not very complex and it can very well be projected in advisory and assurance practices as the advisory focuses on the value part while the assurance focuses more on control parts. The manager highlighted that

If you look at the assurance perspective and also PwC as an organization wants to spread is "TRUST". Trust is really an important word for us and if I look into the categories here, a part of the model is actually serving as a value but also factors like compliance, regulation, explainability, accountability, privacy, and security are all attributed towards trust. What we want to do as PwC is to create trust with these new type of solutions. We would rather sell this model by making it simpler by decreasing the number of phases in a presentation towards the customer as they might not understand everything and have the original model for our internal assessment and analysis.

The manager suggested that this model would work perfectly for the mid-level management on-board & operational folks and the more of the condensed version would be enough to get a seal of approval from the clients. The model can be seen as a complete guide as the management would be interested in a particular phase while the data experts would focus on their respective phases and the compliance team would be more interested towards the deployment and monitoring phase.

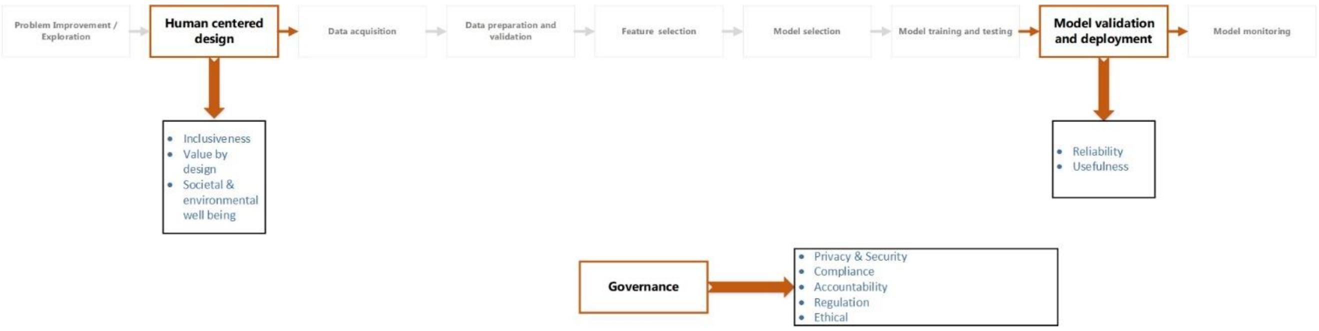


Figure 41 - Evaluation of the model – Insights and recommendation by a manager (Risk Assurance)

ACTOR 3 – MANAGER (RISK ADVISORY) -The third and final interview was with a manager from risk advisory who was also invited to reflect on the model critically and provide his recommendations and insights. The manager has a strong background in IT & Business strategy, IT auditing, and Digital Transformation. He is currently engaged in several digital transformation based solutions and has been in the front line in engaging with the clients.

Same like the previous interview, the manager started his reflections over the model from the human-centered design phase. However, he suggested some touchpoints over the data process as well.

To start with, the manager pointed out that ***value by design*** would be really important for his clients and in the context of AI, as the model or solution is being developed for a certain reason and it needs to have a certain value attached which should be developed right from the initial phase.

Inclusiveness and diversity, though they are important at this phase, it is more to do with the actual dataset itself making sure that they are representative and it would be more of condition that one should have it in place while working on the data sets which can be attributed towards the completeness of the data.

The manager made some interesting insights in the following phases of data and he started analyzing those phase from the auditing perspective and he suggested that factors like ***reliability and accessibility*** would be really important because, at the end, the model has to produce same results over and over while doing audits on the AI systems and it mainly depends on the data that you feed in to the systems. Also the ***completeness*** of the data.

Completeness, in audit terms, is one of the important aspects and it was interpreted with a clear example by the manager,

Let's say that, we have a financial statement report of the client and when we need to audit such reports, we need to make sure that the report is complete and accurate so that, all the information that we expect to be in the report must be there. But, when we need to audit an AI model, we have to validate like did all the inputs make it into the output to determine the completeness of the data.

Consistency of the data can be tied to the reliability as it is believed that if the data is consistent, it can be reliable.

The manager also reflected over the importance of ***responsibility*** as most of the clients really struggle with respect to this aspect in terms of who has the responsibility for AI?

When we implement some normal controls in a business process, the clients usually struggle to pinpoint like who is responsible for the control, who is responsible for checking the controls, who is responsible if anything *goes wrong in the controls. Responsibility is something that is always overlooked but it needs to be set up clearly upfront that there is ownership at each and every phase in this model for instance.*

In the final phase of model validation and deployment, the manager made an interesting observation over AI being ethical.

If you talk about ethical being a factor, it doesn't need to be ethical to work as a model, but it really matters if it is ethical to explain the model to the rest of the world. If you are implementing an AI model in a fast-moving stock trading companies, they probably won't care about AI being ethical and all they would expect is some sort of small profit margins but if you would implement the tool in recruiting, then ethical is a very crucial factor. So, it really depends on the context of the case and the environment.

Regardless of the context and from the auditor's perspective, the manager recommended explainability as the main factor not just towards the models but also the people working on the model. The manager gave an interesting observation over factors like explainability, completeness, and privacy of the model.

As an auditor, you look into the logic of the model, write the report in such a way that somebody else that reads the report should have a clear perception on how the model is working and for this we expect the model to be explainable and only then it would be possible for the auditor to examine such systems. Additionally, the people working on the model have to explain logically how the model works, what has been done to validate the model, etc.

As an auditor, we check for the completeness of the model. One of the main questions that we usually get is "Can you audit this IT system" and they just want to know whether the data in the IT system is complete and accurate and for AI, it would be the same. Is the data in AI is complete and accurate and does the input match the output, etc.

As an auditor, we would look at privacy and security like who can access the model or the data, can that be changed, how secured is the model, etc. Because that will definitely influence the outcome of the model. Privacy, especially in the Netherlands and Europe is gaining high importance especially with GDPR and other regulations coming in. If the AI is handling a large amount of customer data, privacy would be one of the top concerns.

From an auditing perspective, completeness, accuracy, and reliability is very important and is always the case.

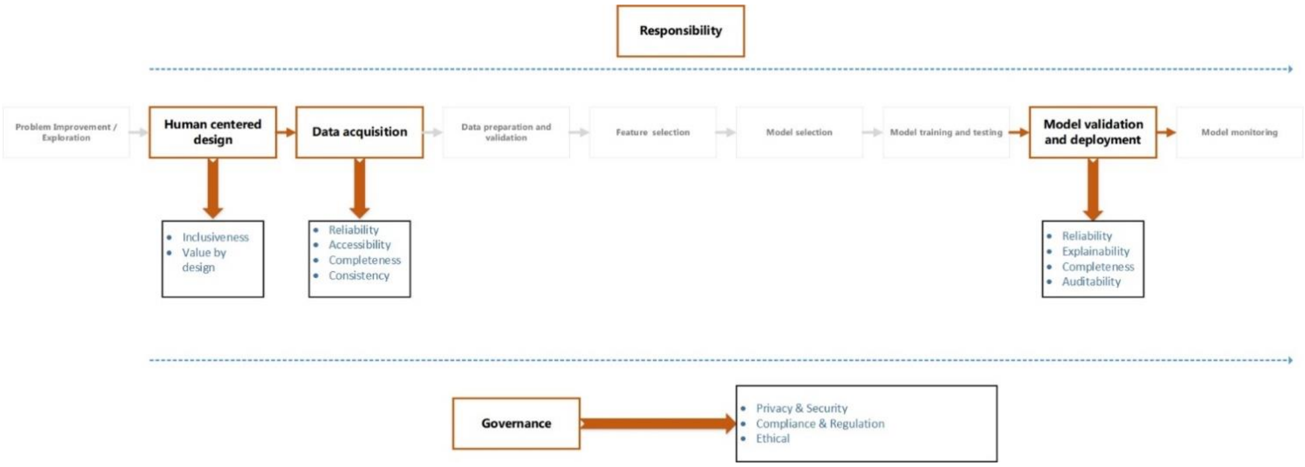


Figure 42 - Evaluation of the model – Insights and recommendation by a manager (Risk Advisory)

EVALUATION OF THE MODEL USING A CASE STUDY

One of the cases was selected to evaluate the model as the case had a typical and full implementation of AI solution. The interview was conducted with one of the relevant data scientists who had been a major part in the implementation of that particular case.

Who was the client? - The client was one of the leading flag carrier airlines operating across Europe.

What was the challenge? – The client's bottleneck was that the spend in their system was not categorized and as a result of this, they had challenges in planning the budget. Very recently, they started working on addressing such issues and see if they can implement any new solutions in order to gain insights from the data. As they had a huge volume of data and doing this manually was going to be time-consuming and expensive through man powers. It is assumed that there are high probabilities of committing errors when humans are allowed to process such a high volume of data manually. So, there was a clear need for using the analytics-based approach in order to handle the client's challenge.

What was the solution provided by PwC? -Machine learning-based approach was applied to scan the database and apply a defined category based on the spending. The data received from the client was first extracted according to the requirements, processed and was made consistent to make sure that the mapping would work. A clustering-based model was then selected to determine the words that match similar invoices.

Throughout the implementation phase, the data scientist was involved and he reflected that factors like accuracy, consistency, and bias-free data were the factors that were mainly considered during the implementation. The data scientist also indicated that security was not seen as an important aspect in this case as the scalability of the solutions was not large and also the data that was used had nothing to do with privacy. Factors like bias-free were not seen in all the phases of the model as it was identified in this research. It was realized that most of the factors that were identified in this research were not considered as the type of solution delivered was basically a clustering-based approach that helps to categorize the themes, words and provide insights.

However, the data scientist was asked if there is any factor that was missing from the model which they actually considered during the case. The data scientist commented that only a few main factors were considered during the engagement and he was able to recognize those factors in this model. The data scientist finally reflected by saying that the *model would really make sense when they start to engage with the clients for building an AI solution that is based on prediction, decision-making or AI-based robot or system deployed in critical environments as it would require intense assessment on these factors that has been described in the model in order to trust such systems*. To summarize the above case, only factors like consistency, reliability, completeness, bias-free, relevancy were considered as important trust factors of data which were seen as DQ dimensions while factors like usefulness, reliability, accuracy, and performance were seen as important trust factors of the AI model. It was very well realized from the case that not all factors identified from the initial version of the model were considered in this case and at the same time, there were no surprising or new factors found from the case and all the factors identified were already a part of the model. The below table gives a detailed portray on the list of factors that were considered during the case.

Phases	Factors
Data acquisition	Consistency, reliability, completeness
Data preparation and validation	Bias free, consistency, relevancy
Feature selection	Relevancy
Model selection	Accuracy
Model training and testing	Usefulness
Model validation and deployment	Reliability
Model monitoring	Performance

Table 21 - Identification of trust factors based on the case analysis

7.3 FINAL VERSION OF TRUSTED AI MODEL

Now by, understanding the recommendation and reflection provided by the targeted participants (data scientist, and managers) over the initial version of the model, the final version of AI trusted model would be developed. The new model won't just contain the factors based on the final perspective provided by the data scientist and the manager during the model evaluation because, it is presumed that there could be different perspective if there had been more respondents interviewed from the same background and by the end it would be appropriate to choose the factors that have the commonality between those respondents over the factors combined with the initial findings from the interview and insights from literature study as that could help to improve the level of generalization. So, the model that would be developed has to ensure some of the key prerequisites that are well established before realizing the value of the model. The prerequisites are here as follows,

Prerequisite 1: The model must enable the identification of the phases that are involved in the development of AI.

Prerequisite 2: The model considers the factors that were identified from the initial literature study, main findings from the initial interview analysis, and finally the insights and reflection provided by the experts over the initial version of the model.

Prerequisite 3: The model mainly emphasizes the trust towards the data and the model in the various phases and focuses on the touchpoints towards the design and the human trust factors as they were realized at the later stage of the research and it is not considered to be the main scope of this research.

Prerequisite 4: The model uses the dimensions of DQ as factors in trusting the data

Prerequisite 5: The model must have the relevant trust factors tied towards each phase of the model.

Prerequisite 6: The model should contain the indicators for each factor that were identified for the respective phases in the background. The detailed indicators for each factor have been presented in Appendix F and G.

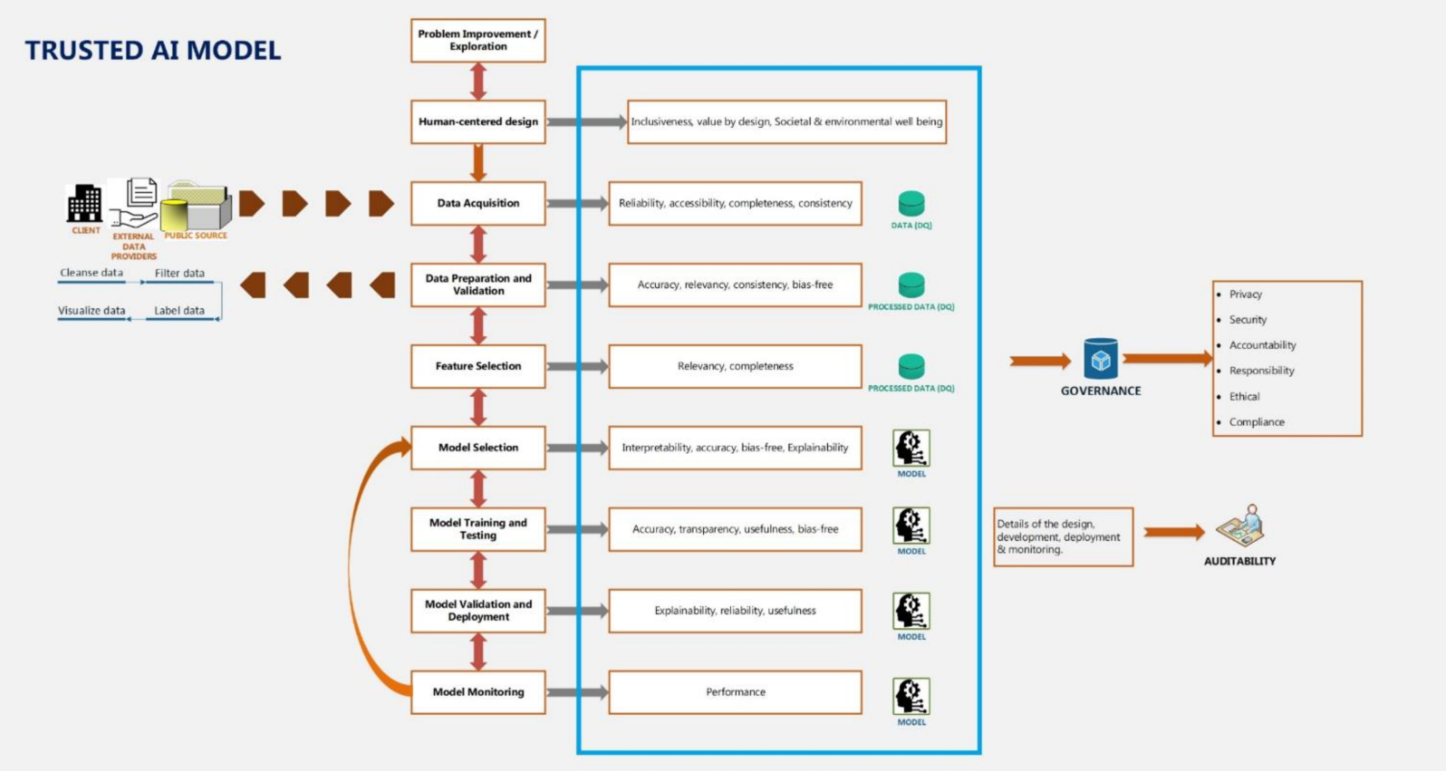

TRUSTED AI MODEL
Problem Improvement / Exploration
Human-centered design
Inclusiveness, value by design, Societal & environmental well being
Data Acquisition
Reliability, accessibility, completeness, consistency
DATA (DQ)
CLIENT
EXTERNAL DATA PROVIDERS
PUBLIC SOURCE
Cleanse data
Filter data
Visualize data
Label data
Data Preparation and Validation
Accuracy, relevancy, consistency, bias-free
PROCESSED DATA (DQ)
Feature Selection
Relevancy, completeness
PROCESSED DATA (DQ)
Model Selection
Interpretability, accuracy, bias-free, Explainability
MODEL
Model Training and Testing
Accuracy, transparency, usefulness, bias-free
MODEL
Model Validation and Deployment
Explainability, reliability, usefulness
MODEL
Model Monitoring
Performance
MODEL
GOVERNANCE
Privacy
Security
Accountability
Responsibility
Ethical
Compliance
Details of the design, development, deployment & monitoring.
AUDITABILITY

Figure 43 - Final version of trusted AI model

7.4 MODEL COMPARISON

With the growing attention that has been paid towards AI to ensure that they are responsible and trustworthy, several technology giants like Google, IBM, Microsoft has established the important themes and principles in order to build trustworthy AI systems. Additionally, several research institutions have been established to mainly address the concerns and the challenges that AI can pose to society and these institutions have introduced their own themes and guidelines. Very recently, a high-level expert group on AI set up by the European Commission had released a document that provides the guidelines for building a trustworthy AI. The main point of introducing these technology giants and other AI research institutions at this point of time is to ensure whether the model has considered and addressed the important themes and are in line with the principles laid by these companies and institutions. This can improve the level of usefulness in the real world context and more importantly, add value to the proposed model and to this research as a whole.

Few things have to be noted in prior is that the research would presume this validation as an additional milestone. Secondly, the themes identified by these groups (AI-based research institution & leading tech companies) are considered as the trust factors in the context of the trusted AI model and it would be ensured whether such themes are present in this model. Third, the model couldn't be compared with any existing models or frameworks as it is assumed that there aren't parsimonious trust models proposed in the scientific research especially in the field of AI and only the trust models in the field of automation, human-robot interaction is found to be abundant. Also, if the proposed model has to be compared with the existing trust models studied in automation, human-robot interaction, then the proposed model won't be fully explored & compared as the dimensions of DQ are also seen as trust factors in the model. In such a case, the proposed model has to be compared separately with existing trust models in the field of automation & other systems and with the existing DQ models studied in different sectors and in the field of big data. This could only bring complexities and it would be hard to determine the real value of the trusted AI model as a whole if the model has to be compared based on the above case. Finally, the section won't cover on each of the themes comprehensively as they were listed and given a short description in chapter 3.7. An outline of all the themes identified by the technology-based companies and research institutions is shown in table 22.

ORGANIZATIONS	THEMES
IBM	• Fairness • Robustness • Explainability • Lineage
Microsoft	• Fairness • Reliability & Safety • Privacy & Security • Inclusiveness • Transparency • Accountability
Google	• Fairness • Interpretability • Privacy • Security
PwC	• Fairness • Interpretability • Robustness & Security • Governance • System ethics
EU Commission	• Human agency and oversight • Technical robustness and safety • Privacy and data governance • Transparency • Diversity, non-discrimination, and fairness • Society and environmental well being

	• Accountability
Partnership on AI	▪ Safety-Critical AI ▪ Fair, Transparent, and Accountable AI ▪ AI, Labor, and the Economy ▪ Collaboration between people and AI systems ▪ Social and Societal Influences of AI ▪ AI and Social Good
AI now Institute	▪ Rights & Liberties ▪ Bias & Inclusion ▪ Labor & Automation ▪ Safety & Critical Infrastructure
The Institute for Ethical AI & Machine Learning	▪ Human Augmentation ▪ Bias Evaluation ▪ Explainability by justification ▪ Practical accuracy ▪ Trust by privacy ▪ Security ▪ Displacement Strategy ▪ Reproducible Operations

Table 22 - Overview of the themes laid by tech companies and AI research institutions

At first sight, it can be clearly noticed that factors like fairness, privacy, security, reliability, etc. have been one of the priorities for the organizations and research institutions. The term "**fairness**" emphasized by several groups can be defined as the ability to treat all people fairly. In the light of AI, Fairness can be related to, having representative data sets, checking the system for unfair biases and addressing them, using training and testing data that are free of biases to avoid the unfair treatment. With respect to the model proposed in this research, fairness can be tied to the factor of bias-free as it has been identified in various phases of the model. For example in the data preparation and validation phase, bias-free is one of the major factors to ensure that the processed data is free of bias and is a representative of the user population. Similarly, in the phase of model training and testing, bias-free is a trust factor to make sure that the training and testing data doesn't have any bias and any potential skews or variance found during training and testing had to be addressed and resolved. Special importance was laid to factors like privacy and security as the system would be using large volumes of data to make a prediction or a decision and it should be ensured that the data & the system as whole are completely safe and secured and at any cost, it is not vulnerable to tampering or comprising the data that is being trained and tested on. In the context of the model, it can be realized that those factors are actually not tied to any of the phases but instead, they have to be monitored throughout the whole stage. It was well understood from the interviews that having governance in place would assess factors like privacy, security and establish accountability, and responsibility clearly right from building the model to the monitoring of the model.

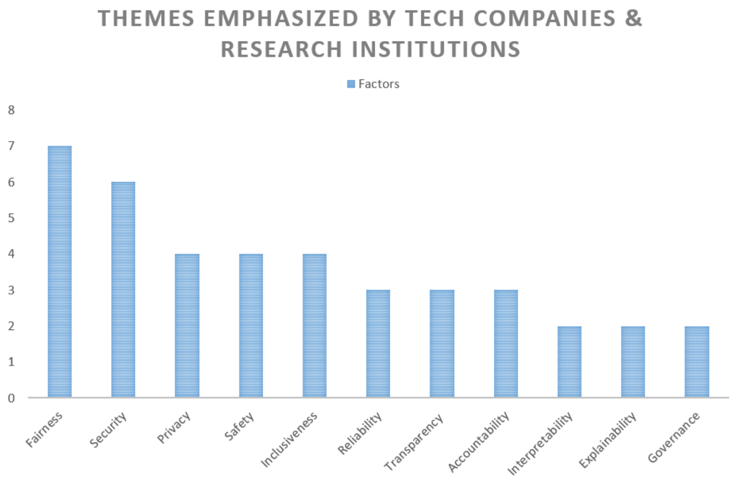


Figure 44 - Prioritization of themes based on the analysis (Literature- Themes of AI)

On top of that, factors like inclusiveness were highly been recognized by the AI research institutions and some of the tech organizations. Inclusiveness highlights about engaging the right set of people in the early design phase, active stakeholder participation, involving a diverse set of representatives, targeted group of the application, industry and functional domain experts, etc. This can be clearly noticed in the human-centered design phase of the proposed model. Furthermore, factors like transparency, interpretability, and explainability were regarded as a critical factor and it has become one of the top priority for every organization to build a model that needs to be explainable, interpretable, and transparent. In a certain way, some of these factors are interlinked. If one can build the model that is explainable, it would easily help the users to anticipate the behavior or functioning of the system (Interpretable). (Oxborough et al., 2018) classified transparency and explainability as the two dimensions to interpretability. Transparency would help to shed light on black-box models, while explainability helps the people to understand the decision making of the AI system. Models that make a decision without a reason or justification or with no traces on how it had arrived at a decision is called black-box models.

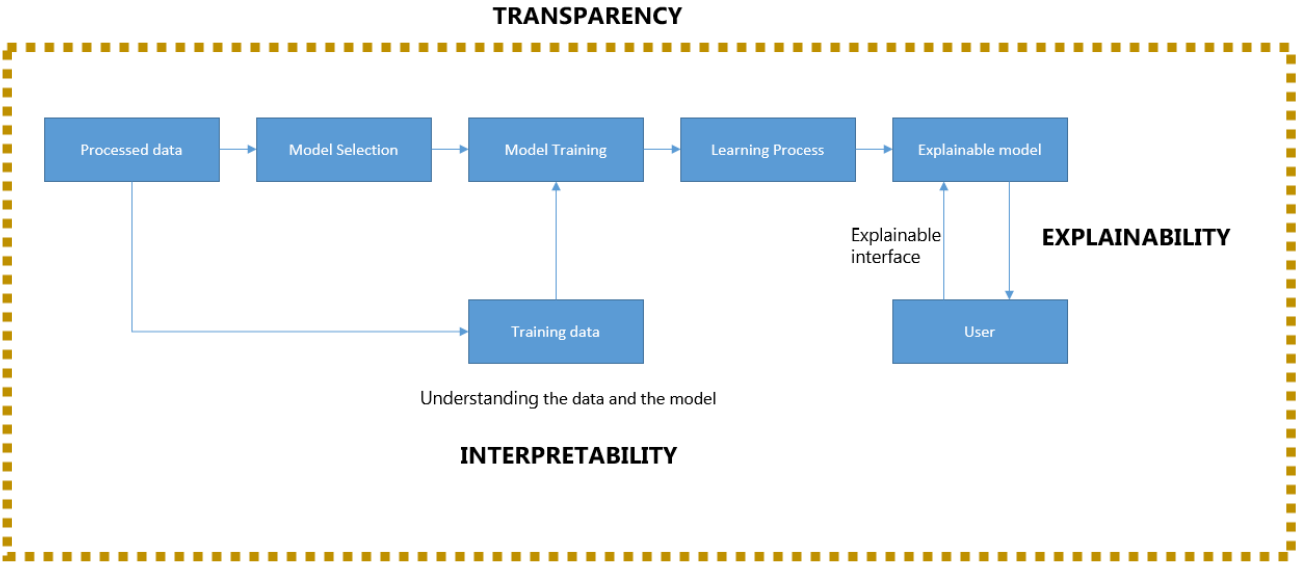
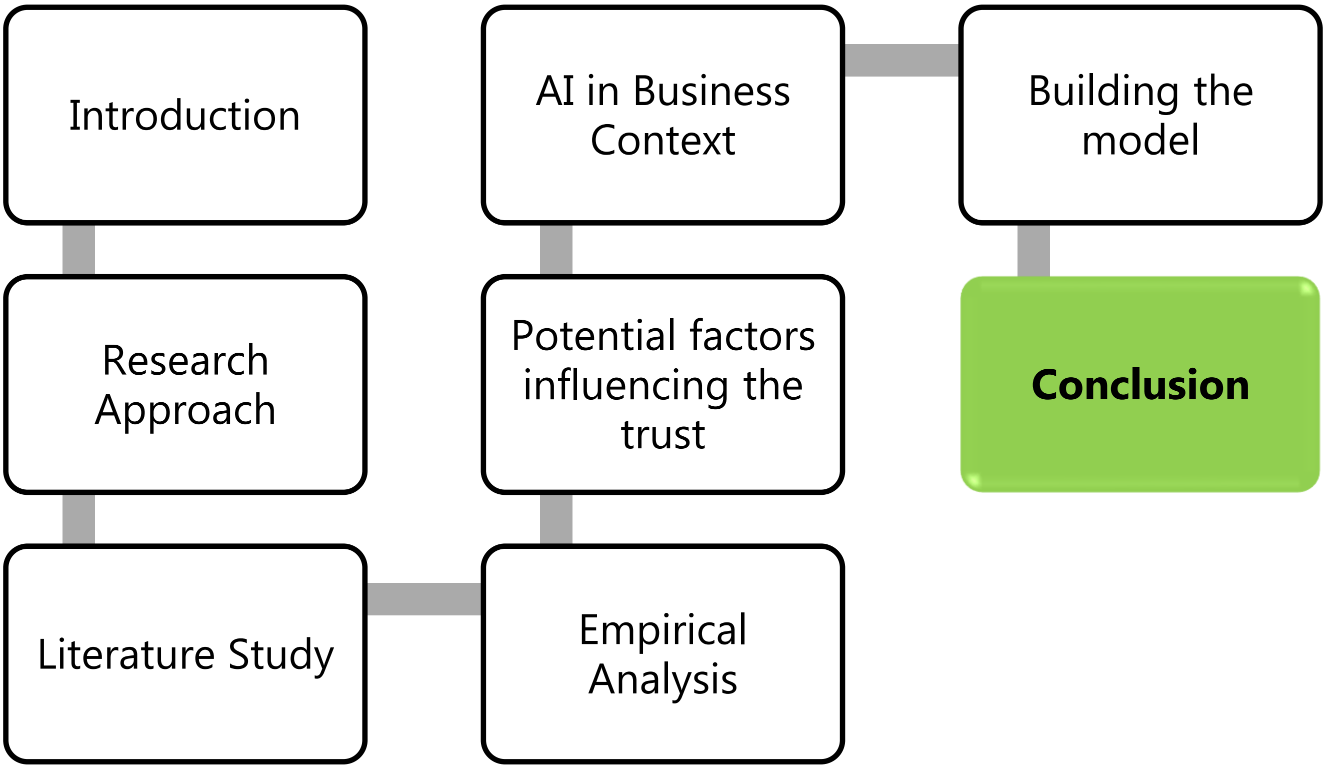


Figure 45 - Illustration of explainability, transparency, and interpretability in the development of AI (Partial stages)

When seen in the context of the model, it has to be ensured whether model chosen can explain the results, whether the information can be interpreted and these factors are tied to the model selection and training phases respectively. When the model is made explainable, interpretable and transparent, it can improve the auditability of such models as at the end, examining the model thoroughly will boost the trust over the technology. To conclude, the model has considered nearly all the factors addressed by AI institutions and companies. Though the themes laid by these institutions can be seen in the broader context, the proposed model clearly projects the importance of every theme & their value which are seen as trust factors and trying to position them to the relevant phases of the model as to, where they would need to be assessed in the development of AI. Most importantly, the proposed model has clearly projected the main dimensions of data quality serving as a trust factor to trust data and thereby, making a clear distinction on the trust factor between the data and the model. These dimensions of DQ are mainly seen in the phases of data acquisition, and data preparation and validation while the trust factors of the AI model are seen from model selection phase to the model monitoring phase. The core themes identified in the proposed model (trusted AI model) can be found in Appendix H.

8

CONCLUSION



Artificial intelligence has always been the buzz over the recent years especially in the business sector. Several leading tech companies and startups had invested in AI by delivering AI-based products and producing data-driven AI-based solutions. These solutions or products are bound to do a specific task but when AI starts getting complex and moves into the stage of enhanced and cognitive automation, the value of trust comes into the major spotlight. Leading research institutions, scientific researchers, and tech companies have already started addressing the importance of trust towards AI. Almost every stakeholder, potential users have major concerns like - **Does AI explain the results? How is AI using my data? Is AI been governed?** etc. and all these concerns address to one main question - **How can I trust AI or Whether AI can be trusted?**. This research has aimed at identifying the important trust factors associated with the data and the resulting AI model based on the literature review, desktop research, and interview with the actors involved in the development of AI and develop a trusted AI model accordingly.

The assumption for this thesis **is that realizing the trust factors associated with AI and assessing them upfront using a model can improve the overall trust of the technology and thereby providing a seal of trust to the stakeholders, potential users, and to the investors of the technology.**

So, the main objective of the thesis is to,

Develop a model to assess and improve trust in artificial intelligence.

As the outcome of this research, a trusted AI model has been developed that includes several phases involved in the development of AI with relevant trust factors associated with each phase of the model. The final version of the trusted AI model can be seen in figure 43 and the detailed indicators for each trust factor are available in Appendix F and G.

The answers to the main research question and sub-research questions would be clearly depicted in chapter 8.1. To realize the major value to the academic research, essential trust factors of data, AI model and the factors englobed around them would be identified in chapter 8.2. Also, reflection over the proposed model would be discussed in chapter 8.3. At this point, all the working hypothesis that was formulated in the earlier phase of the research had to be validated and this is done in chapter 8.4. This would later be followed by the practical and theoretical contribution laid in chapter 8.5. The research would also lay limitations involved in this research, which would be presented in chapter 8.6 and the research would end with the recommendations for future research and PwC that can be seen in chapter 8.7.

8.1 OVERVIEW OF THE RESEARCH QUESTIONS

At first, the answers to sub-research questions that were framed for this research will be answered as this would certainly provide the base and considerable inputs in answering the main research question.

1. What are the various risks involved in AI?

In this research, the risks were categorized into two phases which are namely the data phase and the model phase. In the phase of data, one of the main sources of risk is the data itself and it can be due to bad data, poor data quality, outdated data, incomplete data, etc. and these instances would influence the biases which are a major source of risk. The major elemental aspect for any AI applications are the data and when such data have inherent biases, the resulting outcomes would be prejudiced as well. Most of the controversial incidents that had occurred to date has been due to the biases in the data. Secondly, privacy and security aspects of the data have been an increasing concern in the context of AI as it is expected that AI applications would require a large amount of customer information in order to make an efficient recommendation or personalized suggestions to the user. In such stance, the owners of the data would expect that their data is not being compromised at any cost and therefore would expect transparency on how their data is being handled, in what ways their data is made secure. So, lack of transparency on data process can also pose risks not just to the owners but also to the auditors as they would find it challenging when they are asked to audit such process of AI application. Additionally, there would be a need for proper governance established around the data process. Lack of governance is also one of the potential risk encircled around AI as governance is where a clear line of accountability and responsibility are established and by setting up other data protocols around the data process. These protocols involve who has access to the database, how is the data being handled and protected, whether the data is in compliance with GDPR, etc. In fact, governance actually serves as an outer shield to factors like security, privacy, accountability, responsibility, compliance, and identification of biases in the data. So, when there is weaker governance established, there would be much room for a security breach, data privacy risks, accountability & responsibility risks, etc..

In the phase of the model, one of the major challenges is opening the black-box model. Black-box models are systems or algorithm that makes a decision without an explanation or justification. Business leaders and users highly expect the model to be explainable and interpretable at least to a certain extent. Lack of explainability and interpretability is one of the risks of AI models. Furthermore, auditors would expect the model to be transparent in order to audit the AI systems thoroughly. To thoroughly examine the AI models as a part of auditing, the model is mainly expected to be explainable, interpretable and more importantly, being transparent and this is apparently lacking in the AI applications. So, one can experience the risk of not being able to audit the systems due to the models that are not transparent, explainable and interpretable which are also considered as a major risk in the medium of AI. Lack of governance around the model is also considered to be a major risk. In the context of the model, governance is valued high as that's where regulation, ethical aspects along with privacy, security, accountability, and responsibility plays a major role. Other risks that are encircled around AI would be the performance and reliability risk. If the AI system doesn't produce desired results or outcomes, the model won't be reliable and trusted. Also, the security aspects of the model would also come into the picture as security is not only ensured around data but also towards the model.

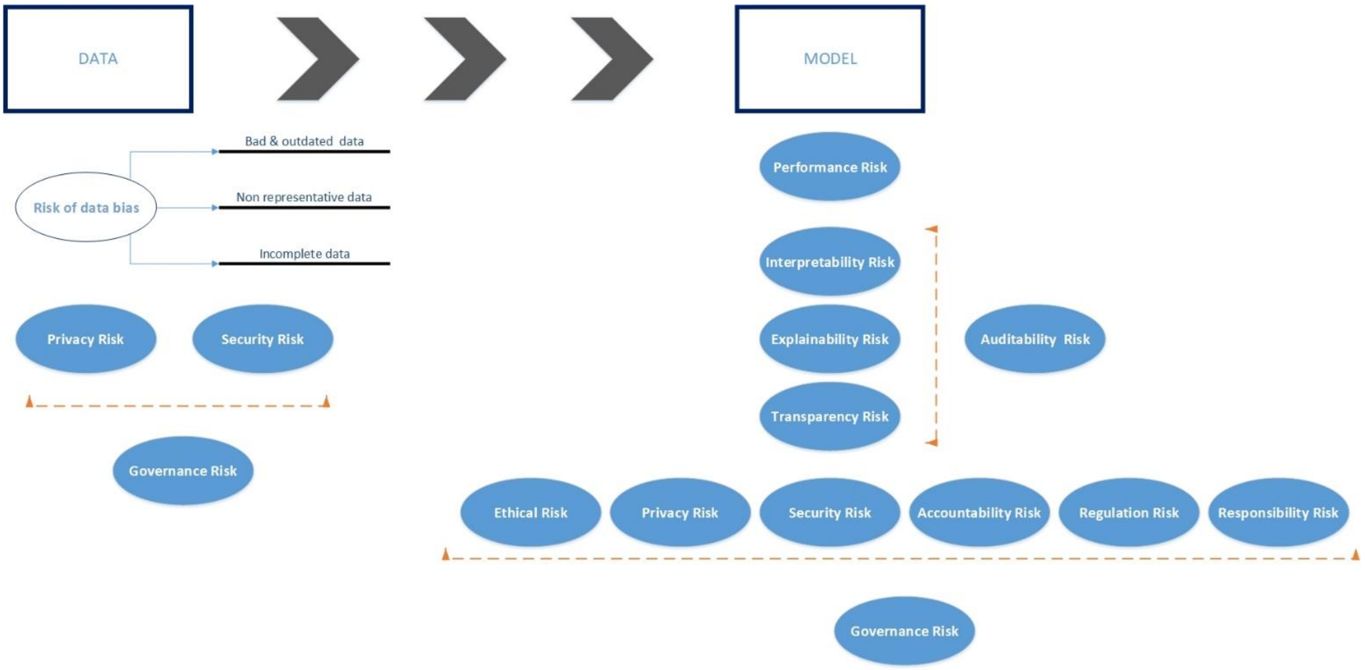


Figure 46 - Various risk around AI

2. *What factors of trust influence the human-machine relationship in the current literature?*

Factors like predictability, dependability, and faith which were seen as important in interpersonal relationships were used as the main reference when the initial study was done on trust towards automation. For example, predictability would form the initial source of trust and it would mainly depend on its systems behavior, assessing system predictability accurately by the humans. Based on this, dependability would form the basis of trust after a certain point of time and it refers to the extent to which the human can rely upon the systems. Faith would be the final phase in trust formation which is formed based on the future behavior of the system. Furthermore, factors like the performance of the system, responsibility, reliability, robustness, and usefulness of the system were some of the common and most frequently referred factors retrieved from the current literature. To be more specific, it was also found that dependability, reliability, level of automation, and transparency would be some of the crucial factors that mainly influences the trust towards the robot from the perspective of a human operator. To summarize, **dependability, predictability, level of automation, reliability, performance, responsibility, robustness, transparency, usefulness were some of the major factors that influence human-machine relationships**. Some of these factors were used as the main reference in the context of AI. Although factors like predictability, dependability, level of automation were initially used as factors in the medium of AI, it had to be discarded at the end, as the actors interviewed dint emphasize those factors. Only the factors such as reliability, transparency, performance, usefulness that were identified as trust factors in human-machine relationships made a greater value in the context of AI. In particular, reliability and usefulness were regarded as very important factors in the surroundings of AI.

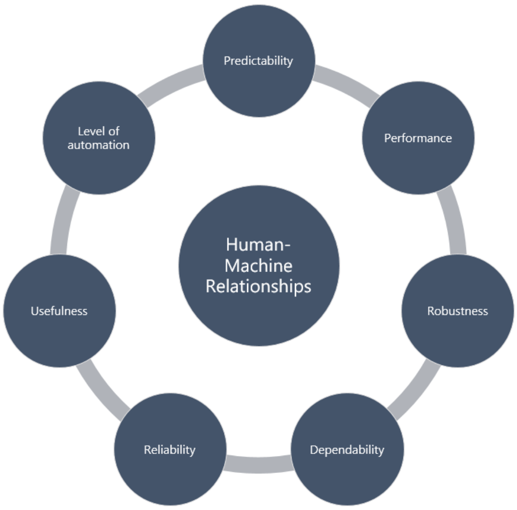


Figure 47 - Trust factors influencing the human-machine relationship

3. *What are the dimensions of DQ that helps in improving the trust in the data?*

Based on the initial findings from the literature review, it was found that dimensions like accuracy, consistency, completeness, availability, accessibility, relevancy, security, interpretability, timeliness, bias-free (objectivity), auditability, reliability, and usefulness of the data were recognized as the potential dimensions of DQ & these dimensions were regarded as trust factors to trust the data and would be more relevant in the context of AI, and big data. However, these dimensions were validated in order to clearly understand its importance from the perspective of actors involved in the development of AI. Based on the analysis and findings from the initial interview, dimensions like accuracy, consistency, bias-free, reliability, and completeness were highly regarded as important dimensions that would influence the trust towards data in the medium of AI. Combining the findings from the interview and initial analysis done on the literature, an initial version of the model was developed comprehensively. When the model was evaluated again by conducting interviews and with the use of a case study, it was noticed that dimensions like completeness, consistency, reliability, and bias-free were the highly recommended dimensions. In addition, dimensions like relevancy & accessibility were recommended too and these dimensions were not seen as crucial from the findings of the initial interview as it could be seen in figure 34 and 35. These DQ dimensions were well positioned in some of the major phases of the model. For instance, reliability, accessibility, completeness, consistency were the main dimensions that need to be considered in the phase of data acquisition, and in the phase of data preparation and validation, dimensions like relevancy, consistency, bias-free, and accuracy were the identified dimensions. So, to clearly determine the dimensions of

DQ in order to trust the data in the surroundings of AI, dimensions such as **reliability, accessibility, completeness, consistency, accuracy, relevancy, and bias-free** had to be assessed.

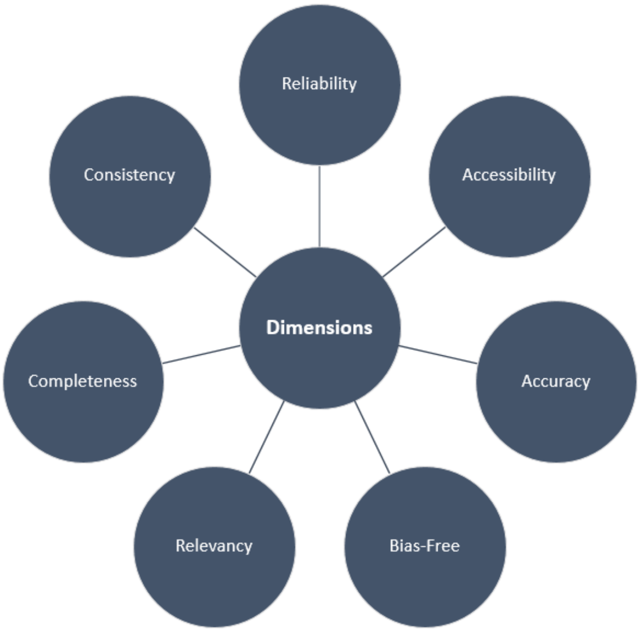


Figure 48 - Dimensions of DQ in the medium of AI

4. What are the core themes of AI?

Various themes and principles have been laid by leading technology giants and some of the renowned AI research institutions in order to build AI systems that are trustworthy and responsible. Every organizations and institution had their own themes and principles and they can be seen in chapter 3.7. It has to be recognized that these themes were assumed as trust factors in this research and it was compared with the final version of the trusted AI model to ensure whether those themes had been addressed in the proposed model to improve the validity and value of the model. The identified themes were ranked based on the repetitions of the themes and it was found that factors like security, privacy, fairness, safety, reliability, transparency, explainability, interpretability, and governance were the main themes. There were increasing attention mainly to fairness, security, privacy, and safety. Next to that were the factors like inclusiveness, reliability, and transparency which were considered equally important. Fairness, in this context actually refers to having representative data sets, using training and testing data that are error-free and bias-free, checking the system for unfair biases. In fact, one could clearly see the importance of fairness which is bias-free with respect to the model, has been paid attention in several phases of the model. Secondly, when AI becomes more and more sophisticated, the business leaders and stakeholders would expect to know the rationale of how the algorithm works, how the algorithm arrived at its decisions, so there has been increasing need on making the AI explainable, interpretable and transparent. This could improve the confidence over the outcomes, and stakeholders trust towards AI as a whole. More importantly, auditors would expect those systems to be explainable and transparent in order to audit such systems. Based on the themes laid by these institutions and the findings from the interview, it could be inferred that factors like **fairness (bias-free), reliability, explainability, interpretability, transparency, and governance ensuring the privacy and security** over the data and model, and establishing the **accountability** and responsibility standards would serve as the main themes of AI.

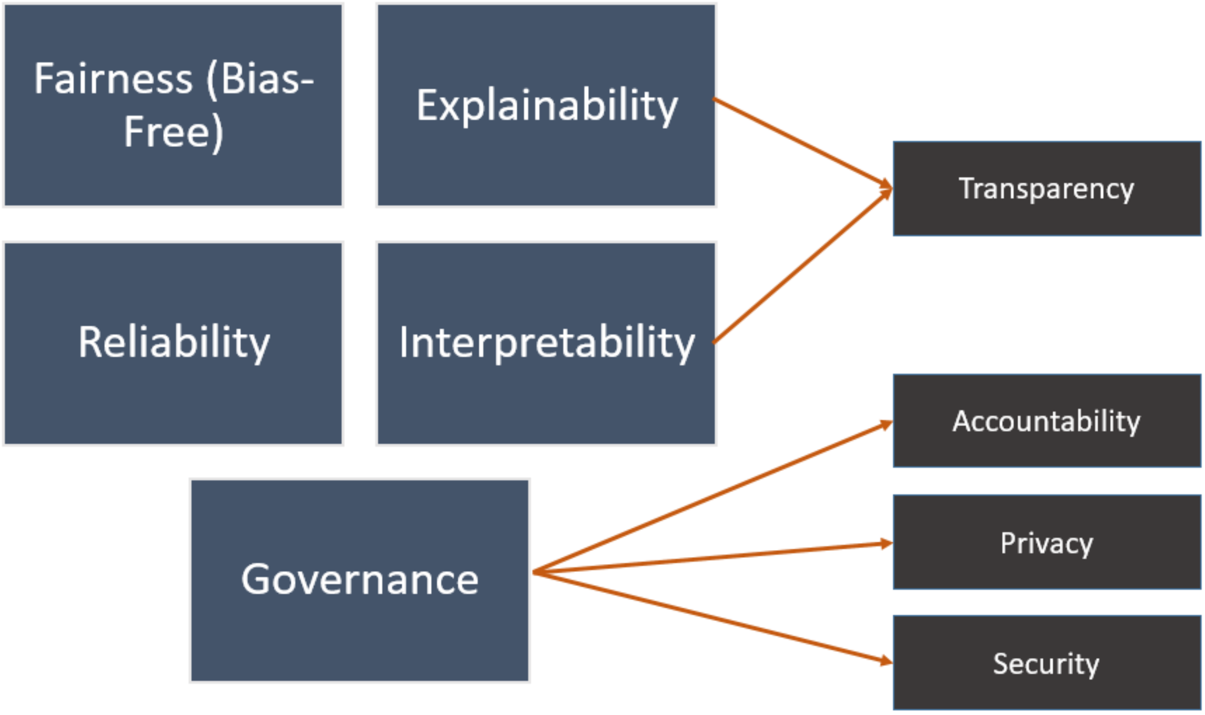


Figure 49 - Themes of AI, Literature analysis

5. How does the model look like that influences the improvement of trust in AI?

The model that has been developed is called as the "Trusted AI model" that includes the detailed phase involved in the development of AI. These phases include problem improvement/exploration, human-centered design, data acquisition, data preparation and validation, feature selection, model selection, model training, and testing, model validation and deployment, and model monitoring. In each phase, relevant trust factors had been identified as these factors had to be assessed in those phases in order to improve the trust in AI. In the phase of data acquisition and data preparation & validation, dimensions of DQ have been used as a trust factor to assess the quality of data and place trust on the data. Special attention was paid in the phase of the human-centered design and subsequent trust factors were also identified. This was realized only at the later stage of the research over the importance of having right stakeholders, diverse people and creating a value upfront in the design phase and this was further strengthened while prioritizing the themes laid by tech companies and research institution. Although the main objective of the research is to focus on the trust factors relevant to the data and model, it was felt that the proposed model would be complete and fitting by focusing on the touchpoints of the design phase and by identifying the relevant factors towards the design phase. The trusted AI model has been visualized in figure 43.

In order to make the trusted AI model complete, the model has to ensure some of the prerequisite upfront and this prerequisite are mentioned below,

Prerequisite1 - Enabling the identification of factors that are involved in the development of AI.

Prerequisite2 - Considering the factors that were identified from the initial literature, findings from the initial interview analysis, reflections and feedback from experts and findings from the case study.

Prerequisite3 – Mainly emphasizing the trust towards the data, and the AI model and focusing on the touchpoints of human trust factor and design factors to make the model complete and fitting.

Perequisite4 – Using the dimensions of DQ as trust factors for the data and ensuring that relevant factor are tied to each phase of the model.

Prerequisite5 – Containing the indicators for each factor identified for the respective phases in the background.

6. What value does the trusted AI model provide?

The development of trusted AI model that visualizes trust factors of the data (in the form of DQ dimensions) & the AI model and projecting those factors clearly in the respective phase of the AI development are one of the key merits òf the model. This can be reckoned as one of the valuable contributions to academic research as the resultant model developed is completely novel & it has emphasized the importance of trust not just towards the AI model but also over the data by studying and identifying the major dimensions of DQ that could play a crucial role in the context of AI. Based on the literature source analysis, there aren't enough model or frameworks proposed, that clearly portrays these trust factors and more importantly, those models or frameworks don't cover the aspects of DQ dimensions.

When seen in the stands of real-world context like what value does this trusted AI model provide to the big firms. Next to the identification of DQ dimensions and trust factors of the AI model, presentation of the trusted AI model itself serves as a main and added value to the organizations. Firstly, the trusted AI model can assist the technology developers (For instance, data scientist, data engineers, developers, etc.) to assess the technology critically using the trust factors identified in the respective phases of the AI development. Second, the top & mid-management levels of the organization comprising of directors, partners, and the managers can use this model as a guide or an assessment tool to validate the probable AI solution at the end or at every phase of the development. The model not only portrays the trust factor present in every phase of the AI development but also provides detailed indicators for every trust factors. Having this indicator along the side can really help the management to validate the technology and see what kind of factors would require more attention when the indicators of respective trust factors don't meet the required expectation. Looking on to the other aspect where there would be a scenario to audit an AI model or application, auditors in such situation can use this model as a guide to examine the technology and solution. Nevertheless, it has to be admitted that the model might not stand as a major value provider in such scenarios but it would help the auditors to be aware of the critical trust factors that need to be assessed while doing the audits of such AI systems.

Furthermore, it was noticed during the interviews and studies done on several white papers that most of the clients who are interested to adopt AI in their business often have concerns or questions with respect to trust and they expect such concerns and questions to be clarified upfront (concerns like Can AI explain the results, Will the technology at any instance produce biased results, Will the technology be reliable enough to trust? etc.). Showing a trusted AI model upfront to the clients can help the clients to have their concerns clarified and would place a substantial level of confidence and trust over the proposed AI solution or application. This would also steer the clients to emphasize more on certain specific trust factors depending on the target environment & context of the case by discussing with management (technology creators). The full potential value of the model would only be attained when the identified trust factors are being taken into account during each phase of the development and ensuring that each factor towards the data and the model are exceedingly meeting the expectations using the indicators as a reference in order fully trust the resulting AI products or solutions. The model at the end would mainly help the management and developers (Technology creators) to establish a robust trust over the AI model or the solutions created and this would encourage them to provide a seal of trust to the investors, clients and other stakeholders involved.

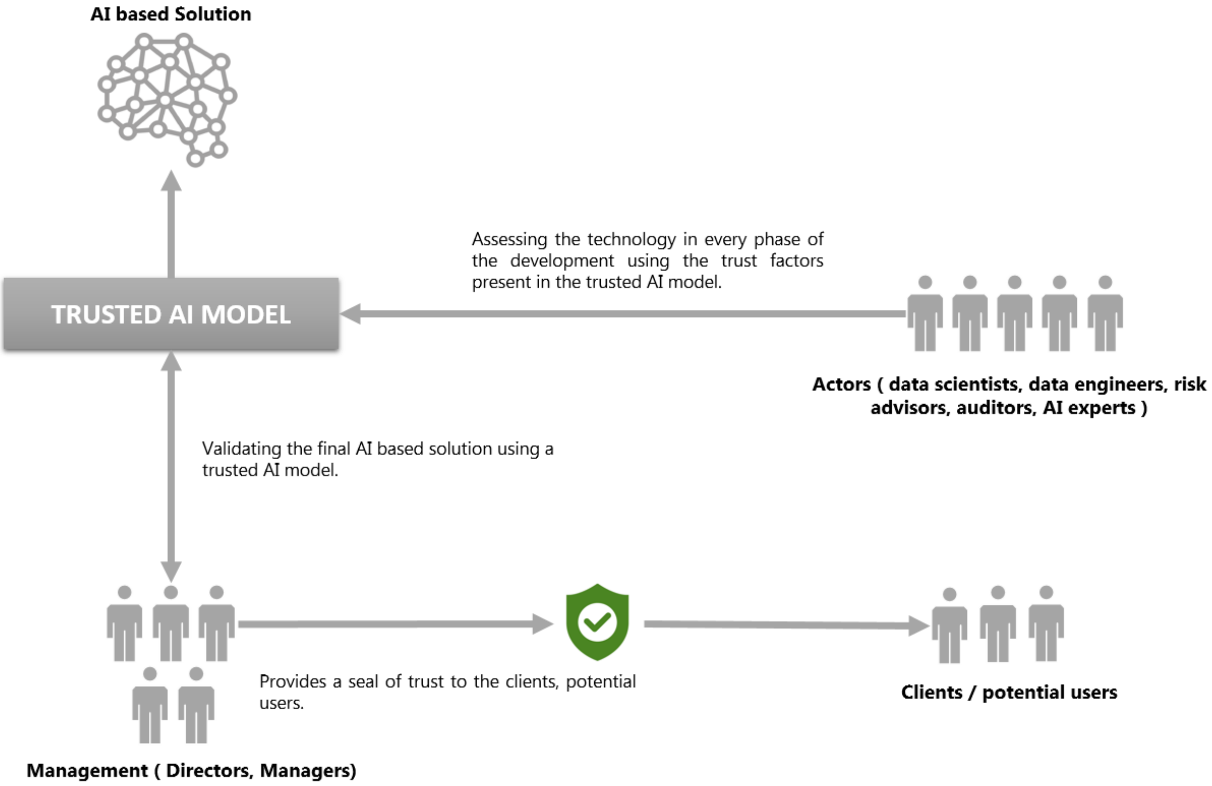

AI based Solution
TRUSTED AI MODEL
Assessing the technology in every phase of the development using the trust factors present in the trusted AI model.
Actors (data scientists, data engineers, risk advisors, auditors, AI experts)
Validating the final AI based solution using a trusted AI model.
Provides a seal of trust to the clients, potential users.
Clients / potential users
Management (Directors, Managers)

Figure 50 - Value of the trusted AI model

Main research question –

What are the factors of trust that influences the trust in Artificial Intelligence from the perspective of actors involved in the development of AI?

Before realizing the trust factors towards AI from the perspective of actors involved in the development of AI, knowing who the actual actors are of paramount importance. Based on how PwC engages with the client for the AI-based solutions, the actors were identified to be directors/ partners, senior managers/managers, data scientist, and data engineer. But, it was further understood that there would be more actors when it involves a large scale AI-based application, say an AI-based robot or software and in such case, one would expect AI experts/specialist, risk advisors, auditors and probably specific industry experts, in addition, to being involved in the development of AI. So, the actors that were interviewed to know the perspective of trust factors towards AI were: directors/partners, senior managers/managers, data scientist, data engineer, AI experts & specialist, risk advisors, and auditors. Findings from the interview with these actors had clearly indicated that factors like **accuracy, auditability, bias-free, consistency, explainability, privacy, reliability, security, transparency, and usefulness** stood out as the important factors as a whole from their perspectives. Upon segregating their perspective clearly towards the data and the model, factors like **reliability, accuracy, consistency, bias-free, completeness, privacy, and security** were regarded as the crucial trust factors of the data while factors like **bias-free, consistency, auditability, privacy, security, reliability, transparency, explainability, etc**. were the identified crucial factor of the model.

By narrowing these factors purely towards the actor specific, actors having a data background like data scientist, data engineer perceived factors like **consistency, accuracy, completeness, bias-free, and reliability** as very important and these factors were seen as the dimensions of DQ to trust the data from their perspective. While the managers from risk advisory and auditors perceive factors like **auditability, security, privacy, reliability, bias-free, transparency, explainability, and usefulness** as the crucial factors and these factors were seen towards trusting the model. It was further realized that governance would serve as core trust factor that needs to be set up around the development of AI and this was highly recommended by most of the actors. It is believed that having **governance** set up would ensure factors like privacy, security, accountability, AI being ethical, and compliance is properly been controlled and governed as these factors at the end really influence the trust in AI. So, governance would act as the main shield over the entire process involved in the development of AI. So, to outline the answer for the main research question. Factors like consistency, accuracy, consistency, bias-free, and reliability were the factors that stood as important from the actors like data scientist, data engineer, and data analytics consultant and these factors were seen as trust factors for trusting the data in terms of the dimensions of DQ, while the factors like auditability, security, privacy, explainability, bias-free, transparency, reliability, and usefulness were seen as trust factors towards the AI model from the perspective of risk auditors and managers from risk advisory practices. A couple of AI experts & specialist who were interviewed regarded **ethical, privacy, and consistency** as the major trust factors of the AI model from their perspective.

8.2 ESSENTIAL FACTORS OF DATA & MODEL

The factors though looks to be comprehensive in paper and in the model, it can be further boiled down to actually determine the essential set of trust factors of the data & the model in the medium of AI. Deriving the essential factors of the data and the AI model would be one of the prime handouts to the scientific research apart from the proposed model. With respect to data, it can be concluded from this study that factors such as **reliability, accuracy, completeness, consistency, bias-free (objectivity), and relevancy** are the key factors that are identified to be indispensable in the context of AI, based on findings from the overall interviews conducted with relevant actors (initial phase & model evaluation) & from the extensive studies done on the literature review and desktop research. These dimensions of DQ are basically considered as trust factors of the data in this context of the research. So, the data that is received and processed has to be **consistent**, **complete, relevant, bias-free, accurate and reliable** and these dimensions can't be ignored and are vital regardless of the context of the case and the target environment. Although dimension like **accessibility** has been stressed in the later phase of the research and has been considered as one of the main dimensions of DQ in the medium of AI, it varies depending upon on the context of the case and in some cases, accessibility won't be seen as the main trust factor. For example, when the clients have their data that is readily available and accessible to the developers (data scientist, data engineers, etc.) and in such scenario, only DQ dimensions such as accuracy, reliability, consistency, bias-free, completeness, and reliability of the data have to be assessed for placing trust over the data and dimension like accessibility won't be a major trust factor in that case.

Now, in the surrounding of the resulting AI model, the most vital trust factors identified from this study are **bias-free, explainability, interpretability, reliability, transparency, and usefulness** of the model. Factors like bias-free, explainability and interpretability, in particular, were highly stressed during the interviews and it was even emphasized highly by several AI research institutions as these factors would ideally help to place a stronger trust over the technology and make AI for good. Most challenges and risk encircled around AI is because of having a biased model or a model that couldn't explain its decisions/recommendations made. It has to be acknowledged that one of the factors, interpretability was not recognized as an important trust factor during the initial interview findings. Most of the actors emphasized the need for having explainable and transparent models and the need for having interpretable models was realized in the later stages (model evaluation). In general, it is often complex to separate interpretability from explainability and transparency as these trust factors are believed to be interlinked.

Furthermore, **governance** was identified as the most essential trust factor that has to be in place around the whole process (mainly from data gathering to monitoring the final model) and within the governance includes key factors like privacy, security. Much interestingly, one of the trust factors, **auditability** happen to gain high importance from the interview findings as it was believed that examining the AI systems thoroughly can positively influence the trust in AI. It was even understood that AI system developed have to be explainable, interpretable and more importantly, transparent as that would improve the feasibility of auditing such systems. **In such a stance, factors like explainability, interpretability, and transparency of the model would be the main**

prerequisites for auditing the AI systems. Perhaps, this was one of the interesting perception that was realized when the risk auditors were interviewed.

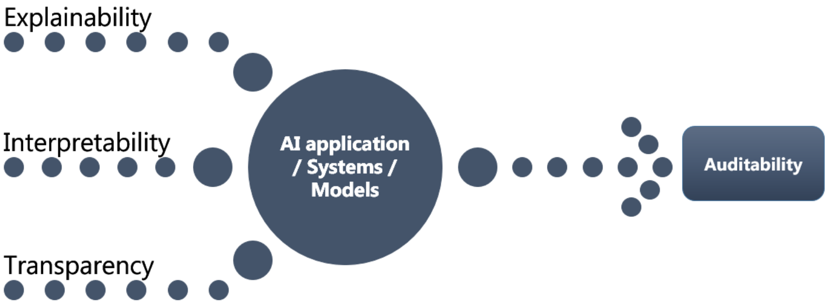


Figure 51 - Prerequisites (trust factors) for auditability of AI systems

In addition, it has to be noted that auditability was not emphasized much by the AI research institutions and tech companies as a key factor and it can be seen in figure 44 where there is no indication of auditability as a key trust factor, and instead, it was inferred by some of the leading tech companies that having a model explainable, transparent, and interpretable would be easy to audit such systems. The below figure 52 depicts the final set of essential trust factor of the data and the AI model and main trust factors englobed around them.

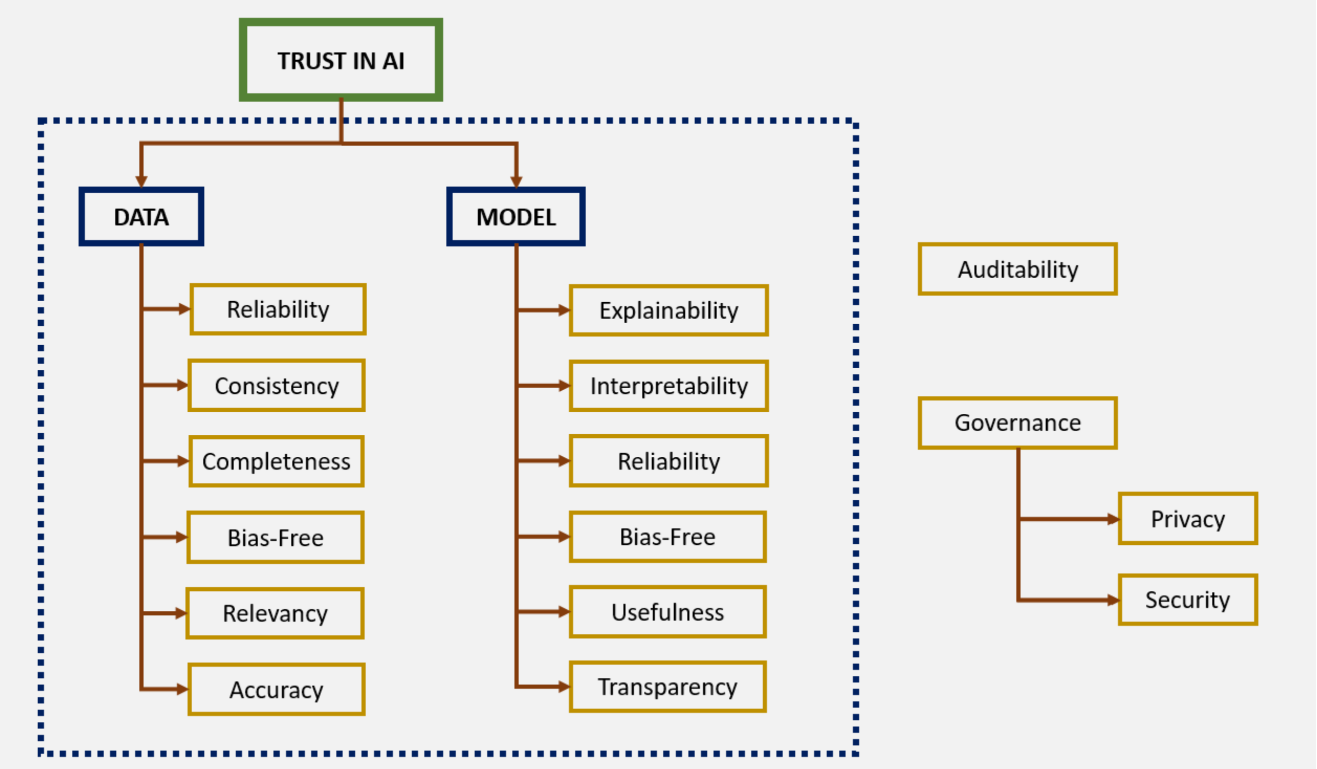


Figure 52 - Essential trust factors of the data and the AI model (Parsimonious model)

8.3 MODEL REFLECTIONS

The proposed model developed would be more appropriate for the large enterprise firms who can utilize the full potential value of the model. The value that the model can provide to these big firms has been elaborated in one of the sub-research questions in chapter 8.1 (Sub-research question 6). The various phases formulated in the model were based on the standard stages used for building one of the dimensions of AI called machine learning proposed by (Guo, 2017). Perhaps, it is believed that these identified phases are more or less in sync with most of the leading tech companies' process flow in developing the AI-based products or solutions and this was further understood while analyzing the white papers and reports published by the big firms. But at this point, it has to be admitted that the model won't be efficient when it is being used by small and medium-sized enterprises (SME) or any AI startups. It is assumed that most of the SMEs and startups have a flexible workflow which is more towards an agile strategy approach and this can be seen predominantly in such enterprises and startups. They don't necessarily follow the same phases identified in the proposed model. Nevertheless, the trust factors which is one of the prime aspects of the model would serve as a main useful purpose to these enterprises (SMEs) and startups more than trusted AI model as a whole. Keeping aware of those trust factors especially the dimensions of DQ and assessing them would guide the SMEs and the startups to establish a stronger trust over the resulting AI products or the solutions created. The same value applies to the auditors in big firms where only the essential trust factors would aid those actors to keep in mind while examining the external AI solutions. Therefore, it can be implied that the whole trusted AI model along with the trust factors & its detailed indicators will hold the key for big firms that are developing an AI-based product or solution. On the other side, only the trust factors would provide value to the SME's, startups, and auditors of the big firms.

So, the trust factors present in the trusted AI model provides the main essence to the firms irrespective of the firm size and the process flow of the development. But it won't be appropriate to generalize or to conclude that the identified trust factors would be effective for all the dimensions of AI (Deep learning, Natural language processing, Expert systems, etc.) especially the trust factors of the AI model as they might be subjected to vary. This would require further research on each of the dimensions to make a clear distinction and stronger generalizations. In the case of the resulting AI model, essential trust factors like explainability, interpretability, transparency, bias-free, reliability, and usefulness would be presumed to be crucial trust factors in all the dimensions of AI but a comprehensive study has to be done on each of the dimensions to confirm the presumptions and make generalizations accordingly. In the context of data, it is assumed that dimensions of DQ would more or less be constant in almost all the dimensions of AI. In fact dimensions like deep learning, natural language processing & its sub-dimensions would require more & more data and if any of these are used in the client cases or for building new AI products or systems, dimensions of DQ like accuracy, reliability, consistency, completeness, bias-free, relevancy would still be at the forefront as trust factor for establishing the trust in the data and this is applicable or can be generalized to any dimensions of AI as data is the main fuel to any AI dimensions. So, it can be inferred that DQ dimensions presented in the trusted AI model would be relevant for most of the dimensions of AI while the trust factors of the AI model can't be generalized to all the dimensions of AI as it would require further analysis & validations before making such strong generalizations.

8.4 VALIDATION OF WORKING HYPOTHESIS & CONCEPTUAL MODEL

Several working hypotheses were formulated during the initial phase of the research upon identifying the potential factors from the literature review and desktop research. A clear distinction was made on the hypothesis laid with respect to the data and the model. These working hypothesis had to be validated now based on the findings from the initial interviews with the relevant actors, insights, and feedback gained during the model evaluation and additionally evaluating the model using the case study.

WORKING HYPOTHESIS	INTERVIEW	CASE STUDY
Having data that is **accurate** will positively influence the trust towards the data and its resulting AI model.	✓	
Having **consistent** data will positively influence the trust towards the data and its resulting AI model.	✓	✓
A data that is **complete** will positively influence the trust towards the data and its resulting AI model.	✓	✓
Data that are highly **accessible** will positively influence the trust towards the data and its resulting AI model.	✓	
Data that is **secured** will positively influence the trust towards the data and its resulting AI model.	✓	
Respecting and securing the **privacy** of the data will positively influence the trust towards the data and the resulting AI model	✓	
Data that is **relevant** to the purpose of the goal will positively influence the trust towards the data and its resulting AI model.	✓	✓
Timeliness of the data will positively influence the trust towards the data and its resulting AI model.		
Data that is **interpretable** will positively influence the trust towards the data and its resulting AI model.	✓	
Data that is **reliable** will positively influence the trust towards the data and its resulting AI model.	✓	✓
Data that is highly **available** will positively influence the trust towards the data and its resulting AI model.		
Auditing the data & its process will positively influence the trust towards the data and its resulting AI model.	✓	
Having data that is **bias-free** will positively influence the trust towards the data and its resulting AI model.	✓	✓
Data that is **useful** will positively influence the trust towards the data and its resulting AI model.	✓	

🟩	Validated working hypothesis & essential factor of data
▨	Invalidated working hypothesis
	Validated working hypothesis

Table 23 - Validation of working hypothesis – Trust factors towards the data

WORKING HYPOTHESIS	INTERVIEW	CASE STUDY
AI systems that are **reliable** will positively influence the trust in AI.	✓	✓
Accurate results produced by an AI system will positively influence the trust in AI.	✓	✓
Auditing the AI system will positively influence the trust in AI.	✓	
AI systems that produce results **without biases** will positively influence the trust in AI.	✓	
AI systems that are **consistent** with the results, predictions, recommendation, etc. will positively influence the trust in AI.	✓	
Having an AI system that is **ethical** will positively influence the trust in AI.	✓	
Establishing the **governance** around the model will positively influence the trust in AI.	✓	

Increase in the **performance** of the AI system will positively influence the trust in AI.	✓	✓
AI systems that respects and safeguards the user **privacy** will positively influence the trust in AI.	✓	
AI systems that are **secured** will positively influence the trust in AI.	✓	
Having an AI system that is **transparent** will positively influence the trust in AI.	✓	
AI systems that can **explain** the decisions made will positively influence the trust in AI.	✓	
Interpretable AI models will positively influence the trust in AI.	✓	
Establishing the **accountability** principle will positively influence the trust in AI.	✓	
AI systems that are **useful** will positively influence the trust in AI.	✓	

🟩	Validated working hypothesis & essential factor of AI model
	Validated working hypothesis

Table 24 - Validation of working hypothesis – Trust factors towards the AI model

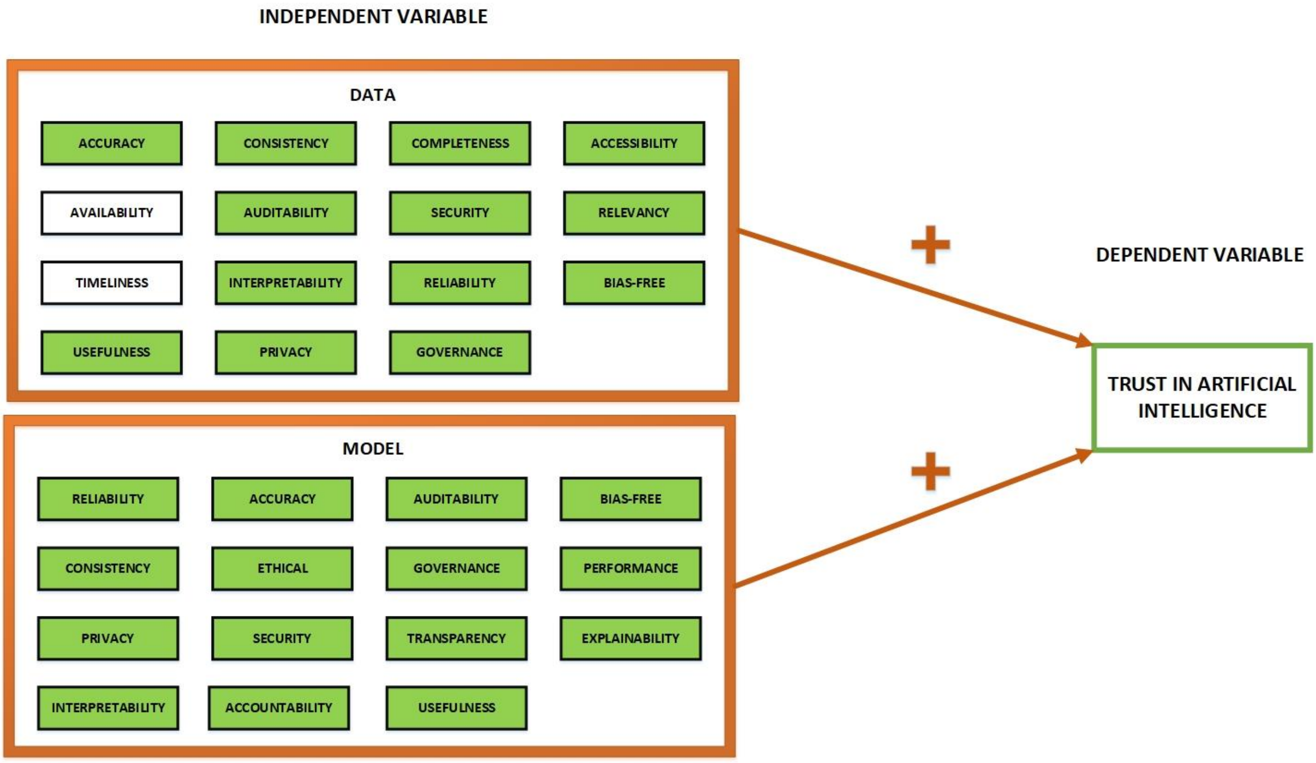


Figure 53 - Validation of the conceptual model

8.5 THEORETICAL CONTRIBUTION & PRACTICAL CONTRIBUTION

THEORETICAL CONTRIBUTION

One of the pivotal aspects of the research is the development of a trusted AI model based on the identification of factors studied from the literature review & desktop research and mainly based on the findings from the interview with the actors involved in the development of AI to clearly understand the important factors influencing the trust in AI. This research contributes to academic literature in various ways and is aimed at filling the possible scientific gaps mentioned in chapter 1.4, that were identified based on the in-depth study of this subject

Firstly, there hasn't been enough research done on the study of trust in the field of AI in particular, from the management point of view as the research has been done from a more technical perspective for improving the trust in AI. This research is considered to be one of the fewer studies that aim to shed light to the scientific community over the importance of trust in AI that is seen from management and socio-technical point of view. One could clearly notice that there is a considerable amount of scientific research done on the study of trust towards automation where several trusted models and frameworks were proposed for improving the trust in automation. But despite the growing attention paid by several scientific researchers and research institutions over the importance of trust towards AI in recent years, there hasn't been adequate models or framework proposed yet for improving trust in AI especially when there is a lot of data and information that are already available. By making use of such relevant information and the factors that were identified while studying trust in automation, the concept of trust & its associated factors were identified with respect to AI. Identifying and studying the factors that influence the trust in AI hasn't been investigated extensively in scientific studies yet and it's a major trend in recent times. This research would tend to improve the awareness of the scientific researchers to emphasize their research more towards building a trustworthy & responsible AI and studying the trust concepts in the field of AI.

One of the major contributions to scientific literature are the dimensions of DQ that has been identified in the medium of AI in this research. There is a vast amount of literature that one could retrieve while looking for DQ dimensions but these dimensions have been identified to play a major role with respect to the certain specific sectors where data is considered as crucial. Very recently, there were studies done on the dimensions of DQ in the context of big data but it was considerably scarce. However, not enough research has been done on the dimensions of DQ in the medium of AI despite understanding the fact that **data is the essential fuel to AI.** This research has identified the relevant dimensions of DQ in order to place trust over the data as it is believed that, improving the trust in data can appease most of the major risks and improve the overall trust in AI. The identified dimensions could open up doors to the experts and other incoming scientific researchers to consider such dimensions and realize what these dimensions have to do even with other emerging technologies like Big data, IoT, etc, and various dimensions of AI. So, this study considers the identification of DQ dimension in the context of AI as one the prime contribution to the academic research.

Building a trustworthy or responsible AI has been the top priority for several technology giants and AI research-based institutions in recent times and these research institutions have addressed some of the factor or themes to build a trustworthy AI. But, the factors laid has to be seen at the broader level and not clearly knowing when would such factors come into the picture and at what stage of the development, these factors have to be addressed in order to trust AI. This study goes one step ahead by clearly identifying the actual phases involved in the development of AI to the readers, and to the scientific community. Second, the relevant trust factors associated with the specific phases have been clearly identified. Third, clearly segregating the trust factors of the data and model rather than generalizing the trust factors as a whole as laid by the research institutions. This complete overview of the model with detailed phases and its associated factor can possibly contribute greater insights and inputs not just for the academic researcher but also to such AI-based research institutions when they start to propose any trusted models or frameworks of AI in future years.

One other added contributions to scientific research are knowing the actual actors involved in the development of AI as most of the research done under the study of AI till date don't emphasize much on who the actors are or the stakeholders involved in the development of AI. This study has initiated by identifying the probable actors who would be involved in the development of AI and of course, it won't be appropriate to conclude that the actors identified in this research are the ones involved in the AI development as the actors could even grow larger when AI becomes more complicated and is still a debatable area. However, the study has tried to throw some light on the scientific community to clearly understand the probable actors involved in AI development and what kind of factors do these actors perceive in order to trust the AI systems. Knowing the trust factors from the perspective of these actors is really vital as they know the ins and outs of the technology like the challenges, risks, and opportunities as at the end, it would be complete only when such perspectives are taken into the consideration and are reflected into the trustworthy AI model.

To improve the value of the research and the resulting model, the study has taken one additional step further by comparing the proposed model with the major themes proposed by some of the leading AI research institutions and tech companies to make sure whether the model has addressed those themes. More importantly, the research has considered the requirements published by the EU commission for building trustworthy AI. Comparing the model with requirements posted by EU commission and other research institutions could improve the validity of the model and such model can serve as a base for the future scientific research as the importance of trust towards AI is expected to grow higher in the coming years.

PRACTICAL CONTRIBUTION

The trusted AI model proposed in this research would be one of the key contributions as to date there is no scientific methodology that has been used to develop a trusted AI model and secondly, there aren't any existing trustworthy models been proposed by tech companies as they have emphasized only the core themes and principles for building a responsible AI. More importantly, there were no indications of how the actual actors in the AI development perceive the trust towards AI like what factors they actually see as important in order to trust AI.

The development of trusted AI model would mainly serve as a tool or a guide for the organizations to critically assess the technology beforehand. This would certainly help the developers of the technology to provide a seal of trust to the investors (clients) so that, they can build confidence over the resulting AI products or solutions and can ideally pass on the same to the consumers when such technology is going to be used by the society. In the end, it won't be the model alone as a guide or tool for the management, but with the detailed indicators for each factor identified in the respective phase of the development. Having this indicator along the side can really help the top management levels (directors, partners), mid-management levels (senior managers/managers) to validate the technology and see what kind of factor would require more actions when the indicators don't meet the required expectation. The model can also serve as a useful purpose when an external organization has developed an AI-based software or a product and it would require auditing on such AI systems or solutions. It might be challenging to realize the full potential of the model but it can serve as a starting point and provide usefulness to the auditors as they would be much aware of the factors that need to be assessed while doing the audits of such AI-based systems.

This study has also made a clear projection on the trust factors towards the data and the model separately and as a result, it could really aid or assist the relevant actors to make sure that such factors had been taken into account and are assessed during the various phases of the model development. For instance, actors like data scientist, data engineers would be more involved in most of the phases in the development of AI and they had to make sure whether the data acquired is complete, consistent, reliable, and accessible to build the initial level of trust over the data and help them move onto the next phase accordingly. Additionally, the research has tried to highlight the importance of the design phase at the later stage of study which stresses on the need of having the right set of people in the early design phase, engaging the right stakeholders, potential users of the application, diverse group of representatives as it is believed that most of the risk encircled around AI seeds from the design phase. So, the organizations have to realize the importance of trust and ensure that they start to build trust right from the design phase by engaging the right experts, all the direct and indirect stakeholders, users of the technology upfront, having a diverse representative and valuing each and everyone's feedbacks before realizing the final design.

8.6 LIMITATIONS

The research did encounter various limitations right from the early planning to execution of the research and these limitations are presented in an orderly way like what kind of challenges did this study encounter.

In the early phase of the research, which is the problem exploration, it was quite clear that there has been greater concern over trusting the technology because of several issues that had come into the spotlight and with technology serving as the black box. This had made the society and business question "whether AI can be trusted". To strengthen the problem statement, there were challenges in finding the relevant scientific literature on trust in the context of AI. So, the problem statement had to be supported with various white papers published by leading tech companies and other prominent AI-based research institutions that had been set up in recent years.

Regarding the research approach, the initial plan was to adopt the case study research methodology as it was assumed that the study would be supported by analyzing at least 4-5 different cases where an organization has delivered AI-based solutions. The main purpose of the research is to develop a trusted AI model based on the factors identified from the literature and findings from the perspective of actors involved in the development of AI in order to clearly understand the important factors. So, it was presumed that using a case study research by analyzing four to five cases would create a valuable study and help in building valid theories from those case study analysis. But, it was rather challenging to find the relevant cases for this research and even when found, not enough information was gathered at the right time. So, the research had to settle with one case and used the literature review & desktop research, findings generated from the interview analysis where several respondents from diverse backgrounds were interviewed to improve the validity of the research and used these methods as the main source.

With respect to the literature review, one of the major pitfalls is the scarcity of scientific papers about the concepts of trust and its factors in the surroundings of AI. It was never easy in finding the relevant factor influencing the trust in AI and it had to be compromised by doing extensive literature analysis over the trust in automation and their trust factors and those factors were considered for the study going with the assumption that automation is one of the early stages of artificial intelligence and it was made sure that the factor selected is really relevant to the context of AI. Secondly, as the value of data is highly valued in this research, there was not enough research done again with the value of data and its associated factors in the AI context. So, dimensions of DQ that were used in various sector and recent research that was done on the antecedents of DQ in the big data context were used as a point of reference to identify the relevant dimensions of DQ in AI context. Nevertheless, it has to be agreed that the analysis was done on certain literature sources & not all sources of literature were utilized which is one of the limitations to this research as well. There could have been papers that would have closely tied to this research and could have added more value by retrieving information & insights from such papers.

Considering interviews as one of the main data collection approach and a big asset to this research, several actors who are and who would be potentially be involved in the development of AI were interviewed. It was hard initially in getting the right data from these actors as it has to be ensured upfront whether the relevant actors fall under the scope of the research. Secondly, delayed waiting time in getting the data from these actors was experienced as the study involves different actors from different backgrounds. Every interview with the actors is a kind of dependency and the study can't proceed further in analyzing until all the relevant actors had been interviewed. For instance, I would have interviewed all the data scientist, data engineer, risk advisors, managers, partners in a time span of first 3 weeks but we have AI experts/specialists as one of the actors in the development of AI, and I was able to schedule the interview with the experts only after 2 weeks. It was later realized that during such waiting times, I could have interviewed more actors from the other backgrounds as that could have added more value to the research. During the analysis of the interview transcripts, there was some extra serious time invested than the time that was allocated during the thesis planning as the text had to be carefully coded and reviewed multiple times because factors like consistency, reliability can be tied to both data and the model and it needs to be clearly seen whether the actor has emphasized the consistency factor to the data or the model. Even a small room for interpretation during the analysis can potentially turn into flawed findings especially when the proposed model is mainly based on most of the findings from the interview analysis.

During the process of building the initial version of the model and its evaluations, it was realized from the interview findings that there was high attention paid on the importance of trust factors over the design phase and the human actors, as those factors could also really influence the trust in AI. The model, however, fails to study those factors extensively as it was realized in the later stage of the research and it had to focus only on the touchpoints over such factors to realize the trusted AI model. The model could have been fully complete if it had considered factors from multi-dimensional aspects but it would rather be challenging to integrate such factor in the model as it would require more validation and it won't be certain to complete the entire research within a time frame of 25 weeks.

Finally, the model has a couple of limitations that could have been still addressed but they have been laid as a recommendation for future research. So, the limitations over the model are: First, the model has identified nine main phases in the typical development of AI but the phases identified were not derived from any scientific source as the main reference in order to prove the standards of the model as there is no single standard process that has been defined in the scientific research with respect to the development of AI. These phases were identified based on the in-depth analysis done on how AI works and this would be in line with most of the organization way of developing AI. Secondly, the model fails to identify the relevant actors in each of those phases as it could really help the management to establish the accountability and responsibility standards which is a crucial factor in the context of AI and this could ideally improve the overall governance and auditability of technology. Third, the model fails to incorporate the perspectives of the end-user which could be the consumer or the client itself as it could have strengthened the robustness of the model. Finally, It has to be agreed that there are no existing models or frameworks that have been proposed in the scientific research explicitly based on the knowledge of the author and as a result, the proposed model could not be compared with the contemporary literature in order to clearly distinguish the difference and the value that the proposed model could possibly provide to that literature. The model could be compared only with themes laid by leading AI research institutions and tech companies.

8.7 RECOMMENDATION FOR FUTURE RESEARCH & PwC

RECOMMENDATION FOR FUTURE RESEARCH

Several limitations that were identified in this research had to be addressed as a part of future research. It can be clearly noticed about the importance and value of trust paid to artificial intelligence in recent years especially from leading tech companies as they face the growing pressure to be answerable to the society, consumers, investors when the technology goes wrong. So, for this, various dedicated research institutions had been set up to address the pressing societal challenge posed by AI and strive to make AI for good. Much surprisingly, there hasn’t been enough or substantial scientific research been done on the trust aspect of AI to date. However, it has started to grab the attention of various researchers and it would be expected that studying and analyzing the concept of trust in the surrounding of AI would be a major trend in the coming years. Researchers have to be encouraged and have to ensure one thing clearly that data is the prime asset to any AI-based applications and their emphasis has to be placed on the data as an initial starting point while studying the trust aspects of AI. It is predicted that the value of data would gain higher importance as the complexity of AI grows bigger & bigger day by day and in such case, researchers have to invest heavily over the importance of the DQ by coming up with detailed framework or model in the field of AI

Secondly, the current study places much importance on studying the trust aspect towards the data and model but it was later realized that trust factors related to human actors and design also plays a major role in influencing the trust in AI. This is ideally opening doors for future research to study those factors extensively in order to realize a robust trustworthy AI model by considering the factors of trust from multidimensional aspects. Since this study mainly involves the identification of trust factors of the data and the model and its entry into the scientific community will be relatively new as such concepts of trust in AI hasn’t been investigated earlier. The scientific researchers and the new incoming researchers had to be encouraged to study those factors more in detail in order to make a clear distinction of the factors between various phases (data, model, human, design). More importantly, factors like ethical, compliance, and regulation would be some of the prime factors as AI becomes the mainstream in the coming years and such factor has to be researched at the broader level. Also until till date, it is quite unclear to understand the actual actors who would be involved in the development of AI. We have been seeing only the recommendations like what kind of actor or stakeholders would be appropriate to be involved in AI applications and of course, it must be agreed that it would be hard to identify the definite set of the actor as it might depend on the context of the application. But still, it can make a clear distinction between the actors who needs to be present regardless of the context of the application (actors like data scientist, data engineers) and the optional actors who would be needed based on the scope of application and what kind of factors do these actors perceive. These questions are trying to shed light for future research as researchers had to analyze the trust concepts from the perspective of actors at a much broader and high level to make a valid generalization.

Also, the research has tried to identify the trust aspects in the context of AI in a general prospect. However, AI has various dimensions like image recognition, text to speech system, robotics, deep learning, natural language processing, and many other applications, etc. This is giving a lot of scope and opportunities for future research to look at the trust aspects of the various dimensions of AI. The researchers have to analyze whether the factors identified from this study would influence the same as when seen at the context of other AI dimensions as it is assumed that the trust factors would subject to change for every dimension of the AI systems. Researching the trust factors in each dimension of AI could really give better clarity on what factors would really mean important to such dimensions and can ideally help in building valuable theories on each of these AI dimensions.

Based on the experience of the author, It has to be agreed that there would be challenges in finding the scientific journal papers on the trust aspects of AI, especially from the management / socio-technical aspects as there isn't enough literature available currently and it is expected to grow in coming years as the importance of trust has already become the main focus in the field of AI. However, in order to make a valid generalization, generate valuable theories from the research, and making the findings more realistic, researchers have to be encouraged to partner with AI-based tech companies and research institutions. Collaborations with such institutions and companies can help the researchers to validate their hypothesis or assumptions made, more accurately by interviewing with the experts or by analyzing the cases of AI-based solutions delivered by those companies. In one way, it can help the organizations and institutions to perceive the value from the research or probably adopt the constructs or models or frameworks proposed from the research and ideally, it would be a win-win situation for all the parties.

RECOMMENDATION FOR PwC

PwC has been one of the organizations that value the importance of trust at the forefront and their main vision is to build and promote trust to their clients and mainly to the society through delivering successful solutions. When it comes to AI, PwC had taken one step further in order to value the importance of trust over AI and ensuring that their AI solutions are responsible and trustworthy. PwC had identified various core dimensions for building a responsible AI and these dimensions include governance, interpretability & explainability, bias & fairness, robustness & security, and system ethics. One of the main value of this research is the development of trusted AI model and this model has actually considered the core dimensions proposed by PwC and is in line with their motive of building a responsible AI. As the model has emphasized the importance of trust factors that need to be assessed in every phase of AI development, it could really help the managers or the directors of the organization to evaluate the technology critically using such a model. While speaking to some of the technology enthusiasts at PwC, it was realized that clients upfront expect some level of confidence or trust to be placed on the proposal if the solution is intended to be an AI-based solution or one of the questions that PwC would encounter from the clients is "How can I trust AI?". In such cases, directors or the managers at the management level should be encouraged to show this model to the clients upfront, and that could possibly boost their confidence. The clients would also be able to understand the process clearly as the model makes a clear perception of the phases involved in an AI development and whether the client concerns like explainability, transparency, privacy, and security are well reflected in the model. So, the managers or directors can consider using such a model upfront during the engagement with the clients. Currently, It has been observed that PwC is engaged in large scale AI-based application and in such stances, this model can really assist the management to facilitate the trust to the clients and other relevant stakeholders.

But before that, it would be highly recommended to try using the model and determine its robustness and usefulness of the model in the real-world context. The model seemed to work perfectly fine when it was evaluated using one of the cases. However, the model has to be evaluated with more completed engagements. This would certainly give insights on how the model is assisting the management and can be altered accordingly. Also, the model would serve a useful purpose for the concerned actors working on the relevant phases in the development of AI as it would help such actors to ensure that the trust factors identified in the specific phase have been met and if not an explanation or justification would be required from such actors.

Finally, the model hasn't covered the importance of responsibility and accountability standards in detail as it has only identified them as a crucial factor for improving trust in AI. If such a model is going to be effective in place for PwC, the actors in every phase have to be identified with the main roles and responsibilities in each of those phases. This would provide more value to the model and increase the awareness & importance of responsibility and accountability principles. So far, we have been seeing this recommendation in the aspect, where PwC has to deliver an AI-based product or a solution. Now looking on to the other side where PwC is one of the leading organization when it comes to auditing. Assuming that PwC would be auditing the external AI applications in the coming years and in such a case, auditors have to be encouraged to use such models or a framework to examine AI applications as that can help them to provide a stamp of trust to the clients. Though one can't realize the full potential of the proposed model, it can at least shed light on the auditors in order to realize the actual trust factors that need to be assessed during the audits of such AI-based systems.

8.8 PERSONAL REFLECTIONS

Over the last ten months, I always hear people talking about Artificial intelligence and the potential benefits that technology can provide. All these buzzes made me learn about the background of AI at first sight. As I started to realize the potential value of this technology by seeing various use cases like Amazon Alexa, iPhone Siri, Google Alpha go, IBM Watson, etc. I got really intrigued and inquisitive about AI and this pushed me to look at the deeper aspect, and that is when I realized the darker side of the technology. One of the major worries for every business leaders and the society right now is "How shall we trust this technology". Undoubtedly, it has to be accepted that AI has a great potential to solve various societal problems and business challenges if trained and implemented correctly but some of the recent controversial incidents and the current challenges posed by the technology has made us reluctant to accept AI. When I first happen to look for scientific papers to see what kinds of factors do actually influence the trust in AI, I did happen to notice that there wasn't much research done on trust aspects of AI from the management or socio-technical point of view. A clear practical problem and a clear knowledge gap from the scientific standards was very well realized at that point, and this motivated me to focus on the identification of trust factors in the context of AI as the main theme in my graduation assignment.

During the initial phase of my thesis journey, I had to look for relevant articles that deals with the trust aspects of AI in general as I couldn't find enough scientific journals explicitly emphasizing on the factors influencing the trust in AI. So, I had to revisit the concept of automation and study extensively on the trust factors towards automation, assuming that automation is also considered to be one of the early dimensions of AI. However, it has to be acknowledged that not all scientific sources were analyzed and there could have been possibilities where there exist some scientific journal on the trust models of AI on other journal sources. Also, in the initial phase, I was able to perceive the importance of data in AI and I decided to value them a lot in this research. I realized at one point that improving the trust over the data, it can influence the trust over the model outcome to a considerable extent as most of the risk entitled around the technology is because of the data that is either bad, incomplete or non-representative. I was taken away by the surprise again when I couldn't find enough scientific papers that emphasize on the dimensions of DQ in the medium of AI despite the fact that data is a core ingredient to AI and where dimensions of DQ would matter a lot. So, I had to first study various dimensions of DQ that was researched based on specific sectors where data was considered crucial and accordingly, identify the dimensions that would be appropriate to consider in the medium of AI. Basically, there was a great deal of time invested in studying the literature and doing desktop research by going back and forth with various concepts and trying to bridge the relation with respect to AI. Though it was very challenging and time-consuming, I personally felt that it was really worth investing so much time on the literature as it actually hinted a perception that,

if I could successfully identify the relevant dimensions of DQ in the context of AI as trust factors towards the data, & relevant trust factors towards the AI model and develop a model accordingly, the entire research would then possibly contribute greater value to the scientific community, shed light to the incoming researchers, and encourage the tech companies to use such model to critically assess the trust factors upfront and improve the trust in AI.

Carrying this motivation thought along, I was able to identify the potential factors influencing the trust in the data and the model based on the comprehensive study done on the literature. Interviews had to be conducted in order to know the important factors from a range of potential factors identified from the literature study. It would only be wise if the relevant actors who are aware of the technology are involved in this study. In such a case, the probable actors involved in AI development had to be interviewed. It did turn out to be very challenging to first identify actual actors in a typical AI development as there is no clear distinction of actors till date. I have to restrict myself by identifying the actors based on how PwC engages with the client in typical AI-based delivery solutions. However, recommendations were laid by some of the AI experts in PwC and those actors were also considered in the study. I initially presumed that I could interview 30+ respondents but eventually, it didn't turn out the way as it was thought as it was very challenging to find a suitable time based on the actor's availability and their interest in the study. I had to finally end up with interviewing 20 respondents but I did try to ensure that all the actors that were identified in this research were represented.

During the analysis phase is where I witnessed one of the pitfalls as I was able to recognize from the findings that most of the actors highly stressed on the importance of trust factors towards the human actors and some actors even highlighted the importance of inclusiveness, diversity as trust factors that need to be considered in the early design phase. These aspects were not elaborated extensively in this research. Only touchpoints were made on those factors in order to make the model complete and fitting. I would have reconsidered & studied such factors in detail only if I had more time to work on this research.

Much on a very personal note - During the early days, I didn't realize the major value that this research would possibly provide as I was assuming that there would be plenty of research done on the trust aspects of AI due to increasing attention in recent times. But when I started digging into the concepts, I was able to figure out the prime loophole and probably that was the striking point that made me realize about the value of this research. When I look back at times where I had spent months and months of time analyzing the literature, talking to some of the technology enthusiasts and AI researchers to validate my initial problem findings, I feel that this study has contributed something new to the scientific research community and much specifically, the dimensions of DQ in the context of AI hasn’t been investigated so far despite realizing the value of data in AI. This is something that I can pat on my shoulder for making me realize the importance of data in the early phase of my research progress and I should be grateful to my supervisors as their insights and recommendations had greatly helped this research, especially while studying on the dimensions of DQ.

I did realize few things which I would have surely considered only if I have got some more months to work on this research and they would probably be, (1) Considering the trust factors of human and design phases in order to make the trusted AI model fully complete, (2) Having realized that there is a clear scientific gap like, identification of factors influencing the trust in AI is missing and identification of DQ dimensions in the medium of AI is missing, I could have rather taken one of the aspects and studied more in-depth as it could have made the research even more sounding and valuable, (3) Though the findings from the interview looks realistic and convincing, I would have interviewed more respondents to make a strong generalization and in fact, I could have made little push in following up with some of the high-level experts on AI who has contributed to ethical guidelines on trustworthy AI published by the EU Commission. The initial plan was to interview one or two experts from the list of contributors but it was rather challenging to find the suitable time to have a conversation with such experts and eventually, it didn't work out. Finally (4), I would have evaluated this model with more number of cases to determine the actual utility of the model in the real-world context.

REFERENCES

2018 AI Predictions: Responsible AI: PwC. (2018). Retrieved from https://www.pwc.com/us/en/advisory-services/assets/ai-predictions-2018-report.pdf

Adams, B. D., Bruyn, L. E., & Houde, S. (2003). *TRUST IN AUTOMATED SYSTEMS LITERATURE REVIEW*. Retrieved from http://cradpdf.drdc-rddc.gc.ca/PDFS/unc13/p520342.pdf

AI for Good Global Summit 2018 Report. (2018). Retrieved from https://2ja3zj1n4vsz2sq9zh82y3wi-wpengine.netdna-ssl.com/wp-content/uploads/2018/12/SDGs-Report.pdf

AI Now Institute. (2017). Retrieved August 7, 2019, from https://ainowinstitute.org/

Awad, M., & Khanna, R. (2015). Machine Learning. In *Efficient Learning Machines* (pp. 1–18). https://doi.org/10.1007/978-1-4302-5990-9_1

Ballou, D. P., & Pazer, H. L. (1985). Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, *31*(2), 150–162. https://doi.org/10.1287/mnsc.31.2.150

Barber, B. (1983). *The logic and limits of trust*. Rutgers University Press.

Batini, C., Di Milano, P., & Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement CINZIA CAPPIELLO CHIARA FRANCALANCI. *ACM Computing Sueveys*, *41*(3), 16. https://doi.org/10.1145/1541880.1541883

Batini, C., & Scannapieco, M. (2016). *Data and Information Quality*. https://doi.org/10.1007/978-3-319-24106-7

Bergstein, B. (2019). Intelligent Machines -This is why AI has yet to reshape most businesses. Retrieved March 4, 2019, from MIT Technology Review website: https://www.technologyreview.com/s/612897/this-is-why-ai-has-yet-to-reshape-most-businesses/

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*(3), 271–286. https://doi.org/10.1016/S0956-5221(97)84644-1

Boillet, J. (2018a). How can you build trust when emerging technologies bring new risks? Retrieved March 2, 2019, from https://www.ey.com/en_gl/digital/how-can-you-build-trust-when-emerging-technologies-bring-new-risks

Boillet, J. (2018b). Why AI is both a risk and a way to manage risk. Retrieved March 2, 2019, from EY website: https://www.ey.com/en_gl/assurance/why-ai-is-both-a-risk-and-a-way-to-manage-risk

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems*, *18*(1), 51–74.

Building Trust in AI and Data Analytics. (2018). Retrieved from https://www.pwc.com/sg/en/publications/assets/building-trust-ai-data-analytics-122018.pdf

Burkhardt, M. (2019). How-To Build Trust in Artificial Intelligence Solutions. Retrieved March 2, 2019, from https://medium.com/omdena/how-to-build-trust-in-artificial-intelligence-solutions-a6d3c7ddf4c3

Butler, J. K. (1991). Toward Understanding and Measuring Conditions of Trust: Evolution of a Conditions of Trust Invenotry. In *Journal of Management* (Vol. 17). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.469.2423&rep=rep1&type=pdf

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(2), 1–10. https://doi.org/10.5334/dsj-2015-002

Cook, J., & Wall, T. (1980). New work attitude measures of trust, organizational commitment and personal need non-fulfilment. *Journal of Occupational Psychology*, *53*(1), 39–52. https://doi.org/10.1111/j.2044-8325.1980.tb00005.x

Corea, F. (2018). AI Knowledge Map: How To Classify AI Technologies. Retrieved June 5, 2019, from https://www.forbes.com/sites/cognitiveworld/2018/08/22/ai-knowledge-map-how-to-classify-ai-technologies/#68da93ac7773

Côrte-Real, N., Oliveira, T., & Ruivo, P. (2017). Assessing business value of Big Data Analytics in European firms. *Journal of Business Research*, *70*, 379–390. https://doi.org/10.1016/J.JBUSRES.2016.08.011

Craglia, M., Annoni, A., Benczur, P., Bertoldi, P., Delipetrev, P., De Prato, G., … Vesnic, A. L. (2018). Artificial Intelligence: A European Perspective. In *European Union*. https://doi.org/10.2760/11251

Crawford, K., West, S. M., & Whittaker, M. (2019). *Discriminating Systems: Gender, Race and Power in AI*. Retrieved from AI Now Institute website: https://ainowinstitute.org/discriminatingsystems.pdf

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women - Reuters. Retrieved June 6, 2019, from Reuters website: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Davenport, T. H. (2018). Can We Solve AI's 'Trust Problem'? Retrieved June 6, 2019, from MIT Sloan Management Review website: https://sloanreview.mit.edu/article/can-we-solve-ais-trust-problem/

Davis, J. (2018). Unlock the Value: From Data Quality to Artificial Intelligence - InformationWeek. Retrieved June 17, 2019, from https://www.informationweek.com/big-data/ai-machine-learning/unlock-the-value-from-data-quality-to-artificial-intelligence/a/d-id/1331076

Dignum, V. (2018). The ART of AI Design —Accountability, Responsibility, Transparency. Retrieved June 19, 2019, from http://designforvalues.tudelft.nl/2018/the-art-of-ai-accountability-responsibility-transparency/

Duursma, J. (2017). The risks of Artificial Intelligence. Retrieved from https://www.jarnoduursma.nl/the-risks-of-artificial-intelligence/

Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *Academy of Management Review*, *14*(4), 532–550. https://doi.org/10.5465/amr.1989.4308385

European Commission. (2019). *ETHICS GUIDELINES FOR TRUSTWORTHY AI*.

Ezry, R., & Tyler, B. (2019). Is your data ready for AI? Part 1. Retrieved from IBM Big Data & Analytics Hub website: https://www.ibmbigdatahub.com/blog/your-data-ready-ai-part-1

FRA. (2019). Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. *FRA – EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS*.

French, B., Duenser, A., & Heathcote, A. (2018). Trust in Automation - A Literature Review. *CSIRO Report*. Retrieved from https://www.semanticscholar.org/paper/Trust-in-Automation-A-Literature-Review-French-Dünser/92f07d3d1356307decb6e97382ad884d0f62668d

Gabarro, J. J. (1978). The development of trust, influence and expectations. *Interpersonal Behavior : Communication and Understanding in Relationships*, 290–303. Retrieved from https://ci.nii.ac.jp/naid/20001538304/

Garbhe, S. (2017). What is Artificial Intelligence (AI) – Becoming Human: Artificial Intelligence Magazine. Retrieved June 11, 2019, from https://becominghuman.ai/what-is-artificial-intelligence-ai-4bde325e5462

Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological Bulletin*, *68*(2), 104–120. https://doi.org/10.1037/h0024833

Guidotti, R., Monreale, A., & Pedreschi, D. (2019). The AI Black Box Explanation Problem. Retrieved June 19, 2019, from https://www.kdnuggets.com/2019/03/ai-black-box-explanation-problem.html

Guo, Y. (2017). The 7 Steps of Machine Learning – Towards Data Science. Retrieved June 21, 2019, from https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hart, K. M., Capps, H. R., Cangemi, J. P., & Caillouet, L. M. (1986). Exploring organizational trust and its multiple dimensions: A case study of General Motors. *Organization Development Journal*, *4*(2), 31–39.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffmann, H., & Söllner, M. (2014). Incorporating behavioral trust theory into system development for ubiquitous applications. *Personal and Ubiquitous Computing*, *18*(1), 117–128. https://doi.org/10.1007/s00779-012-0631-1

IBM. (2018). *Beyond the hype : A guide to understanding and successfully implementing artificial intelligence within your business*. 20.

Janssen, M., Haryadi, A. F., Hulstijn, J., Wahyudi, A., & Van Der Voort, H. (2017). Antecedents of big data quality: An empirical examination in financial service organizations. *In: Proceedings of the IEEE International Conference on Big Data, Big Data 2016*, 116–121. https://doi.org/10.1109/BigData.2016.7840595

Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P. (2003). *Fundamentals of Data Warehouses*. https://doi.org/10.1007/978-3-662-05153-5

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kelly, C., Boardman, M., Goillau, P., & Jeannot, E. (2003). Guidelines for trust in future ATM systems: A literature review. *EUROPEAN AIR TRAFFIC MANAGEMENT PROGRAMME*, (November), 52. Retrieved from http://www.hf.faa.gov/hfportalnew/docsPages/DocPage.aspx?id=1129

Kenett, R., & Shmueli, G. (2016). Dimensions of Information Quality and InfoQ Assessment. *Information Quality: The Potential of Data and Analytics to Generate Knowledge*, (October 2018), 31–52. https://doi.org/10.1002/9781118890622.ch3

Knight, S., & Burn, J. (2005). Developing a Framework for Assessing Information Quality on the World Wide Web Introduction – The Big Picture What Is Information Quality ? *Informing Science: International Journal of an Emerging Transdiscipline*, *8*, 159.

Krogue, K. (2017). Artificial Intelligence Is Here To Stay, But Consumer Trust Is A Must for AI in Business. Retrieved June 6, 2019, from Forbes website: https://www.forbes.com/sites/kenkrogue/2017/09/11/artificial-intelligence-is-here-to-stay-but-consumer-trust-is-a-must-for-ai-in-business/#3248625e776e

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lee, J., & Moray, N. (1994). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information and Management*, *40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Lewiki and Bunker. (1995). Trust in relationships : A model of development and decline . *Bunker, B.B. and Rubin, J.Z.(Eds), Conflict, Cooperation & Justice*, 132–173.

Ma, Y., & Siau, K. L. (2018). Artificial Intelligence Impacts on Higher Education. *Association for Information Systems*, (May). Retrieved from http://aisel.aisnet.org/mwais2018http://aisel.aisnet.org/mwais2018/42

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems. *Decision Support Systems*, *15*, 251–266. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.826.5567&rep=rep1&type=pdf

Marr, B. (2019). Why Every Company Needs An Artificial Intelligence (AI) Strategy For 2019. *Forbes*. Retrieved from https://www.forbes.com/sites/bernardmarr/2019/03/21/why-every-company-needs-an-artificial-intelligence-ai-strategy-for-2019/#7b2bfe568ea9

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709. https://doi.org/10.2307/258792

Mayo, M. (2018). Frameworks for Approaching the Machine Learning Process. Retrieved June 21, 2019, from https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html

McCarthy, J., Minsky, L. M., Rochester, N., & Shannon, E. C. (1995). *A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE*. Retrieved from http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html

Michals, E. (2019). 10 Big AI Failures of the Year 2018. Retrieved June 5, 2019, from Todays in Tech website: https://todaysintech.com/10-big-ai-failures/3/

Microsoft. (2018). *Responsible bots: 10 guidelines for developers of conversational AI*. Retrieved from https://www.microsoft.com/en-us/research/publication/responsible-bots/

Microsoft AI principles. (2019). Retrieved from Microsoft website: https://www.microsoft.com/en-us/ai/our-approach-to-ai

Miller, H. (1996). The multiple dimensions of information quality. *Information Systems Management*, *13*(2), 79–82. https://doi.org/10.1080/10580539608906992

Mojsilovic, A. (2018). Factsheets for AI Services. Retrieved June 21, 2019, from https://www.ibm.com/blogs/research/2018/08/factsheets-ai/

Moorman, C., Deshpandé, R., & Zaltman, G. (1993). Factors Affecting Trust in Market Research Relationships. *Source: Journal of Marketing*, *57*(1), 81–101. Retrieved from https://faculty.fuqua.duke.edu/~moorman/Publications/JM1993.pdf

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence* (pp. 555–572). https://doi.org/10.1007/978-3-319-26485-1_33

Nelson, R. R., Todd, P. A., & Wixom, B. H. (2005). Antecedents of Information and System Quality: An Empirical Examination within the Context of Data Warehousing. *Journal of Management*, *21*(4), 199–235.

Nyhan, R. C. (2000). Changing the Paradigm : Trust and its role in Public Sector Organizations. *The American Review of Public Administration*, *30*(1), 87–109. https://doi.org/10.1177/02750740022064560

Oxborough, C., Rao, A., Cameron, E., & Westermann, C. (2018). *Explainable AI : Driving Business value through greater understanding*. Retrieved from https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf

Peters, M. (2018). Artificial Intelligence, What's New? | Digital Single Market. Retrieved March 4, 2019, from European Commission website: https://ec.europa.eu/digital-single-market/en/blogposts/artificial-intelligence-whats-new

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *COMMUNICATIONS OF THE ACM*, *45*, 211–218. Retrieved from http://web.mit.edu/tdqm/www/tdqmpub/PipinoLeeWangCACMApr02.pdf

PwC Advisory N.V. (2018). Digital Transformation Solutions. Retrieved June 4, 2019, from PwC website: https://www.pwc.nl/en/topics/digital/digital-transformation.html

Redman, T. C., & C., T. (1992). *Data quality : management and technology*. Retrieved from https://dl.acm.org/citation.cfm?id=133848

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in Close Relationships. In *Journal of Personality and Social Psychology* (Vol. 49). Retrieved from https://pdfs.semanticscholar.org/4727/fcf320e6f8c3a8bbd9d7bac22708825f48ad.pdf

Responsible AI Practices. (2019). Retrieved from Google website: https://ai.google/responsibilities/responsible-ai-practices/?category=general

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust1. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Rouse, M. (2018). What is AI governance? - Definition from WhatIs.com. Retrieved June 19, 2019, from https://searchenterpriseai.techtarget.com/definition/AI-governance

Sarker, S., Xiao, X., & Beaulieu, T. (2013). Qualitative studies in information systems: A critical review and some guiding principles. *MIS Quarterly: Management Information Systems*, *37*(4), iii–xviii.

Sekaran, U., & Bougie, R. (2016). *Research Methods for Business*. John Wiley & Sons Ltd.

Shaikh, R. (2018). Feature Selection Techniques in Machine Learning with Python. Retrieved June 21, 2019, from https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics Supply Chain Management View project. *Cutter Business Technology Journal*, *31*, 47–53. Retrieved from www.cutter.com

Sinnott, N. (2018). How Machine Learning Is Changing the World -- and Your Everyday Life. Retrieved March 8, 2019, from https://www.entrepreneur.com/article/312016

Tan, H. H., & Tan, C. S. (2000). Toward the differentiation of trust in supervisor and trust in organization. *Genetic, Social, and General Psychology Monographs*, *126*(2), 241–260. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10846623

The Institute for Ethical AI & Machine Learning. (2018). Retrieved August 7, 2019, from https://ethical.institute/

The Partnership on AI. (2018). Retrieved August 7, 2019, from https://www.partnershiponai.org/

Verschuren, P., & Doorewaard, H. (2010). Designing a Research Project. In *Eleven International Publishing*. Eleven International Publishing.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. https://doi.org/10.1145/240455.240479

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. In *Source: Journal of Management Information Systems* (Vol. 12). Retrieved from http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf

Wang, W., & Siau, K. (2018). Ethical and Moral Issues with AI. *Twenty-Fourth Americas Conference on Information Systems*, (September). Retrieved from https://www.researchgate.net/publication/325934375_Ethical_and_Moral_Issues_with_AI

White, S. A. (2006). *Introduction to BPMN*. Retrieved from https://www.omg.org/bpmn/Documents/OMG_BPMN_Tutorial.pdf

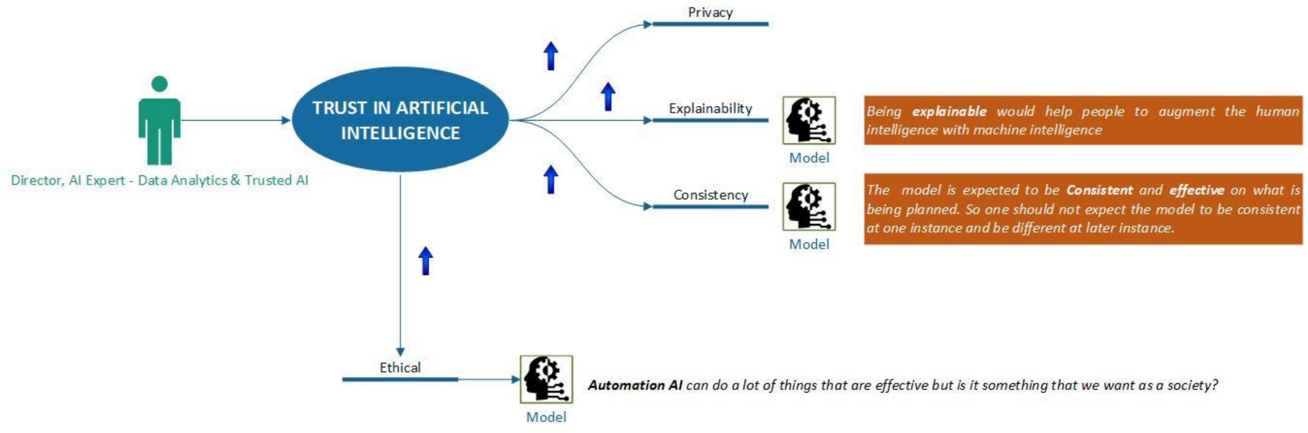
Willems, K. (2017). Data Scientist vs Data Engineer (article) - DataCamp. Retrieved June 21, 2019, from https://www.datacamp.com/community/blog/data-scientist-vs-data-engineer

Woolley, I. (2019). AI and data security: a help or a hindrance? Retrieved June 16, 2019, from https://www.information-age.com/ai-and-data-security-123481333/

Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. New York : Basic Books.

APPENDIX

APPENDIX A - ACTORS PERSPECTIVE ON TRUST FACTORS TOWARDS AI - INTERVIEW ANALYSIS

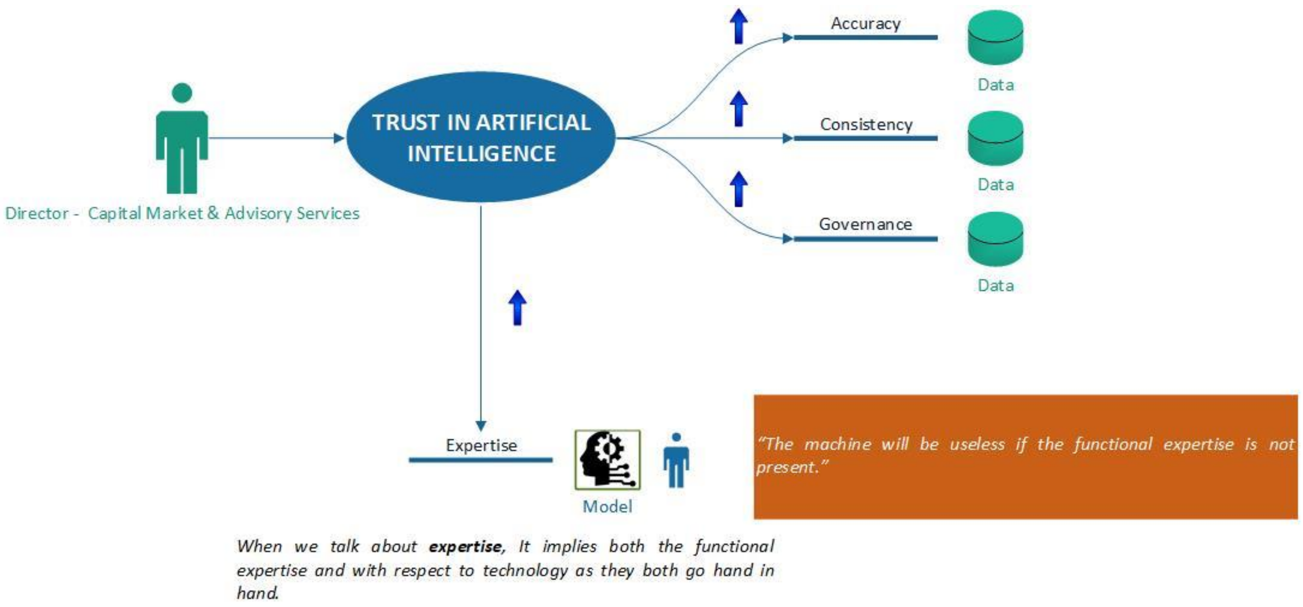

Director, AI Expert - Data Analytics & Trusted AI
TRUST IN ARTIFICIAL INTELLIGENCE
Privacy
Explainability
Consistency
Ethical
Model
Model
Model

*Being **explainable** would help people to augment the human intelligence with machine intelligence*

*The model is expected to be **Consistent** and **effective** on what is being planned. So one should not expect the model to be consistent at one instance and be different at later instance.*

***Automation AI** can do a lot of things that are effective but is it something that we want as a society?*

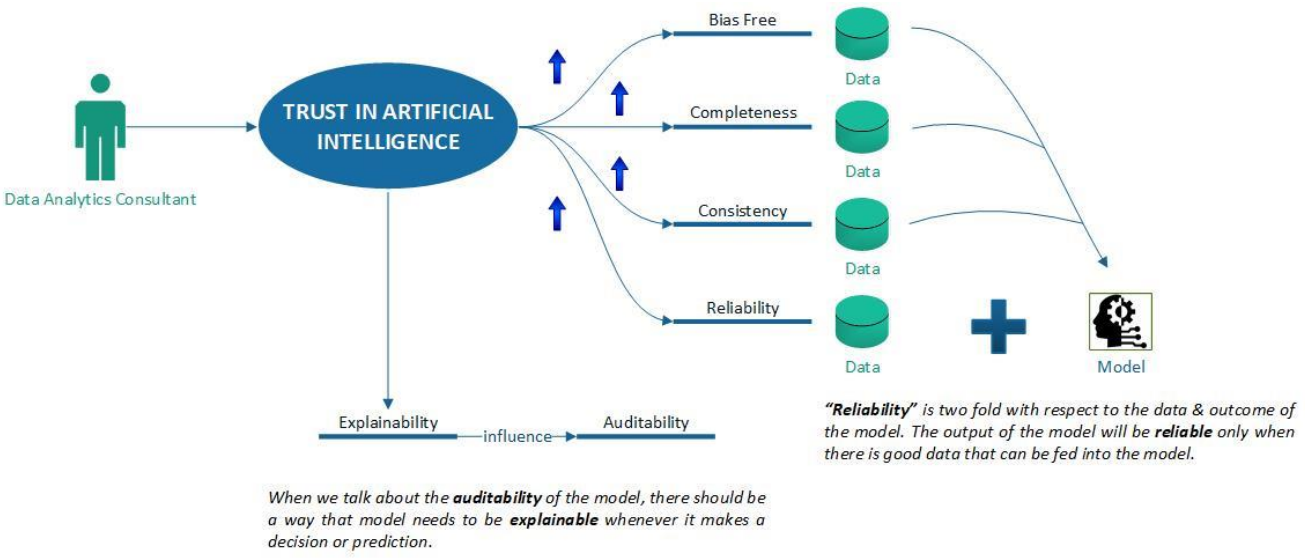
The perspective of a director on trust factors towards AI – Respondent 1


Director - Capital Market & Advisory Services
TRUST IN ARTIFICIAL INTELLIGENCE
Accuracy
Consistency
Governance
Data
Data
Data
Expertise
Model

"The machine will be useless if the functional expertise is not present."

*When we talk about **expertise**, It implies both the functional expertise and with respect to technology as they both go hand in hand.*

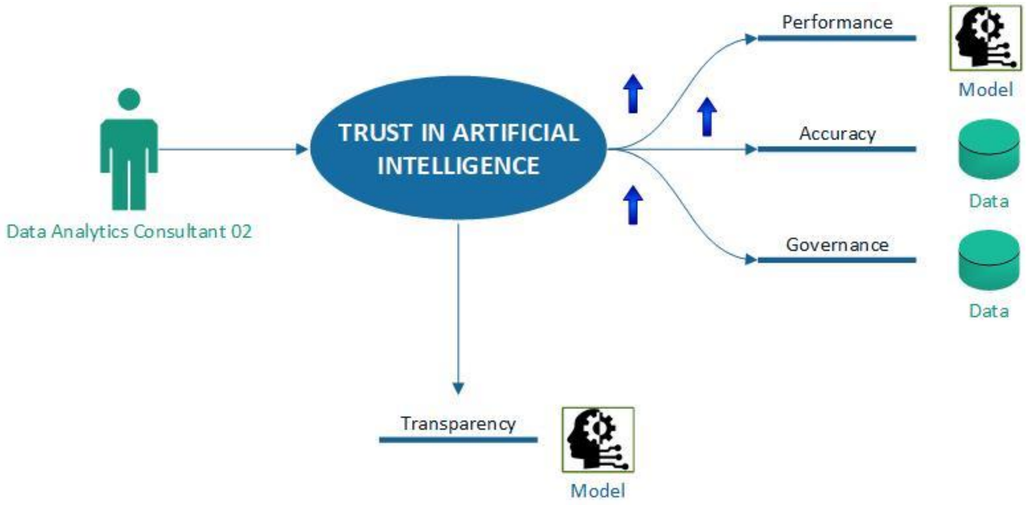
The perspective of a director on trust factors towards AI – Respondent 2



*"**Reliability**" is two fold with respect to the data & outcome of the model. The output of the model will be **reliable** only when there is good data that can be fed into the model.*

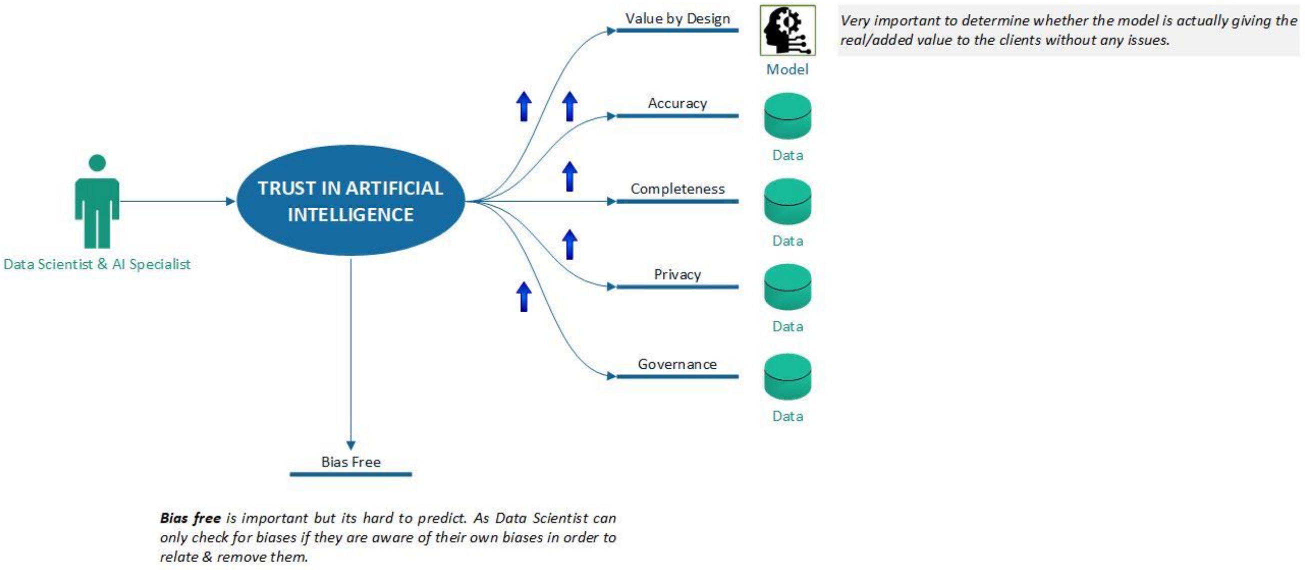
*When we talk about the **auditability** of the model, there should be a way that model needs to be **explainable** whenever it makes a decision or prediction.*

The perspective of a data analytics consultant on trust factors towards AI – Respondent 14


Data Analytics Consultant 02
TRUST IN ARTIFICIAL INTELLIGENCE
Performance
Accuracy
Governance
Model
Data
Data
Transparency
Model

Transparency for example, wont be emphasized at high level. But if you consider any models that has big social impact to the society, transparency is more important in that case.

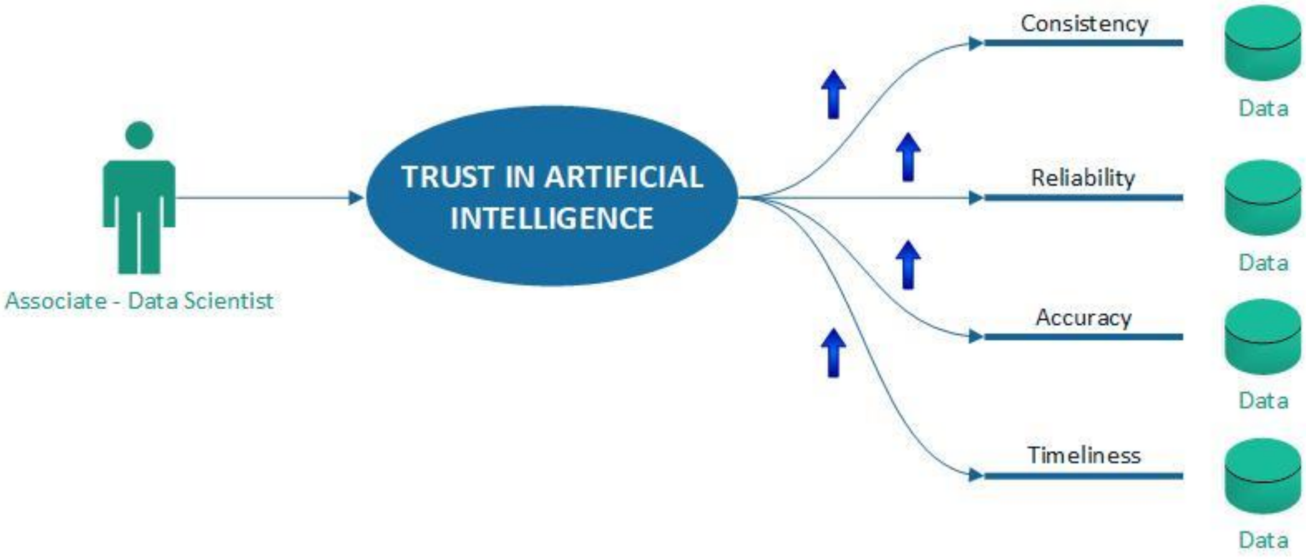
The Perspective of a data analytics consultant on trust factors towards AI – Respondent 15



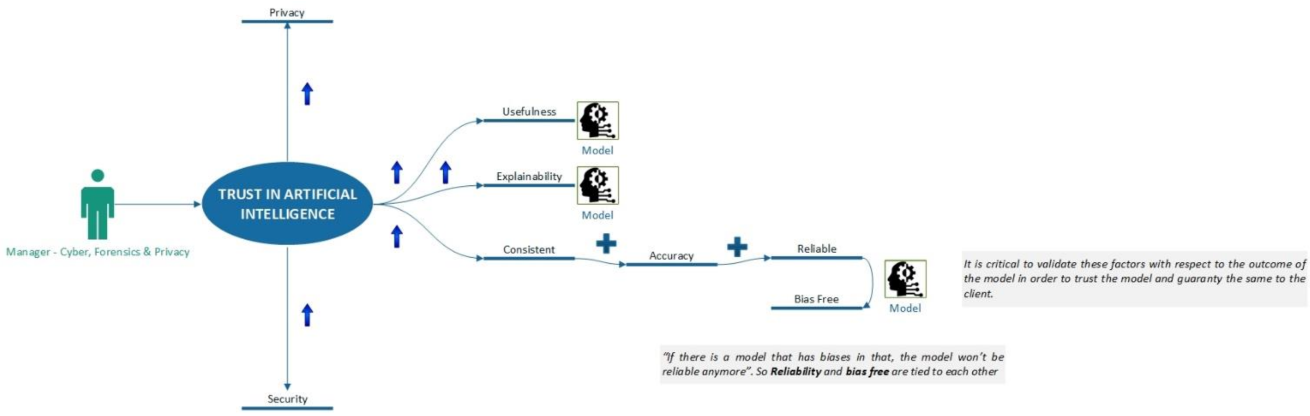
Very important to determine whether the model is actually giving the real/added value to the clients without any issues.

Bias free *is important but its hard to predict. As Data Scientist can only check for biases if they are aware of their own biases in order to relate & remove them.*

The perspective of a data scientist on trust factors towards AI – Respondent 12



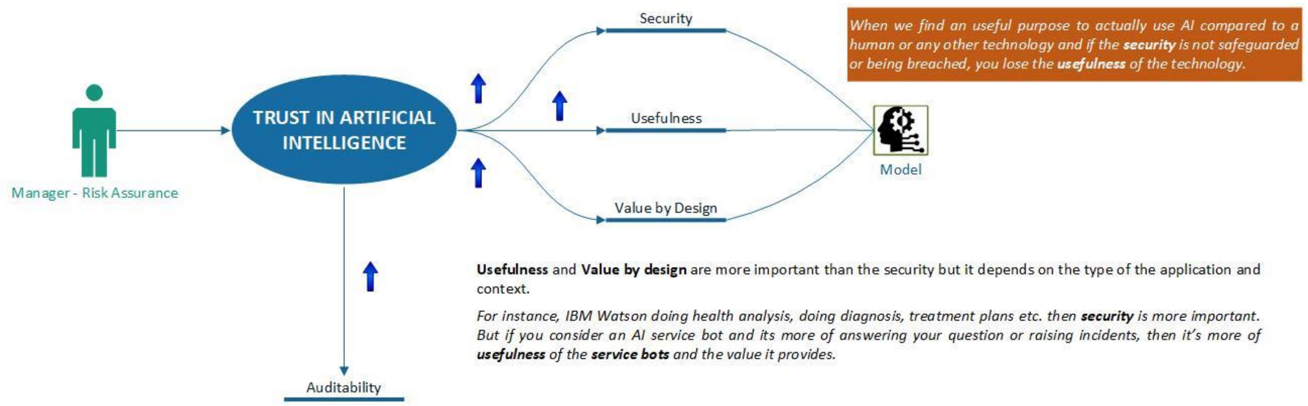
The perspective of a data scientist on trust factors towards AI – Respondent 13



It is critical to validate these factors with respect to the outcome of the model in order to trust the model and guaranty the same to the client.

"If there is a model that has biases in that, the model won't be reliable anymore". So ***Reliability*** *and* ***bias free*** *are tied to each other*

The perspective of a manager on trust factors towards AI – Respondent 4

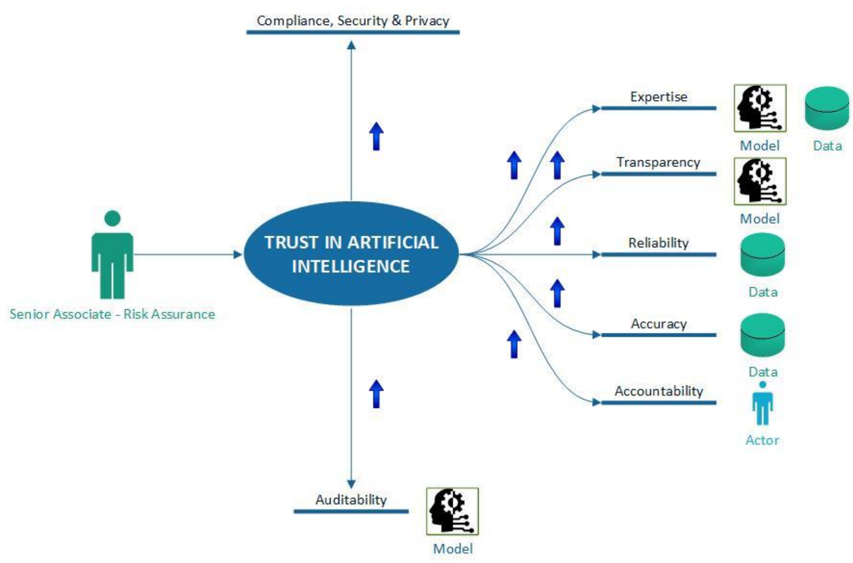


When we find an useful purpose to actually use AI compared to a human or any other technology and if the ***security*** *is not safeguarded or being breached, you lose the* ***usefulness*** *of the technology.*

Usefulness and **Value by design** are more important than the security but it depends on the type of the application and context.

For instance, IBM Watson doing health analysis, doing diagnosis, treatment plans etc. then ***security*** *is more important. But if you consider an AI service bot and its more of answering your question or raising incidents, then it's more of* ***usefulness*** *of the* ***service bots*** *and the value it provides.*

The perspective of a manager on trust factors towards AI – Respondent 7



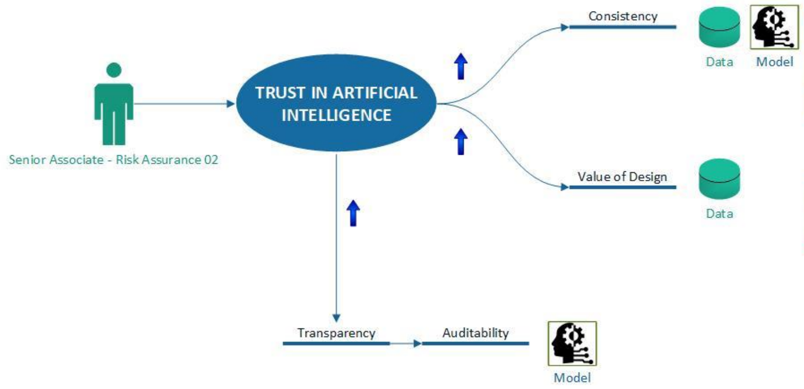
If people are managing algorithm who are incompetent, then it's a risk from the perspective of risk assurance.

If people are working with data that gets fed into the AI are incompetent or haven't been trained on particular data set with respect to the specific context of the AI solution, then it's a risk again from the risk assurance perspective.

If one thinks from the consumer point of view, they are not thinking about the data reliability or accuracy. But they would jump on to the things like the level of autonomy and the performance of AI and such factors can naturally evolve either way.

It more about getting the data aright and having an accurate process.

The perspective of a risk auditor on trust factors towards AI – Respondent 8


Consistency
Data
Model
TRUST IN ARTIFICIAL INTELLIGENCE
Senior Associate - Risk Assurance 02
Value of Design
Data
Transparency
Auditability
Model

There should be a **control** framework over the entire process of AI like not only on building the algorithm but also the on the incoming data, the data that we feed in and finally with respect to the outcomes.

So basically the ***consistency*** of the entire flow right from feeding the data to the algorithm until the outcomes that the machine is producing.

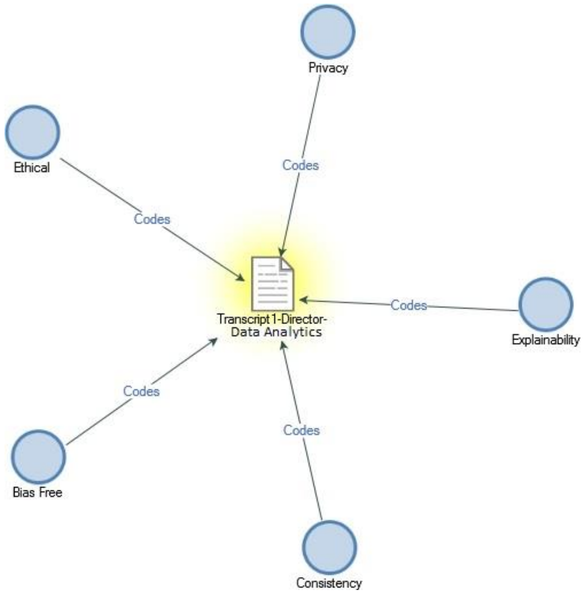
If the machines becomes better & better and we have no control and we will be dependent on these machine and at some point, we can't even head back.

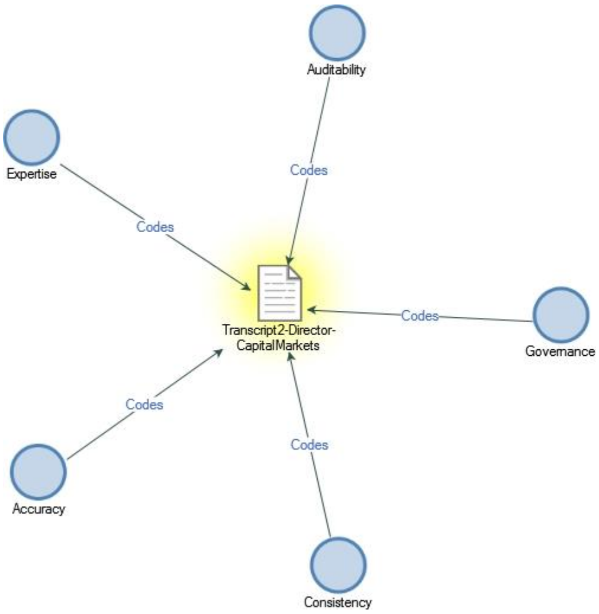
Being ***Dependable*** over the machines is really a worrying factor!

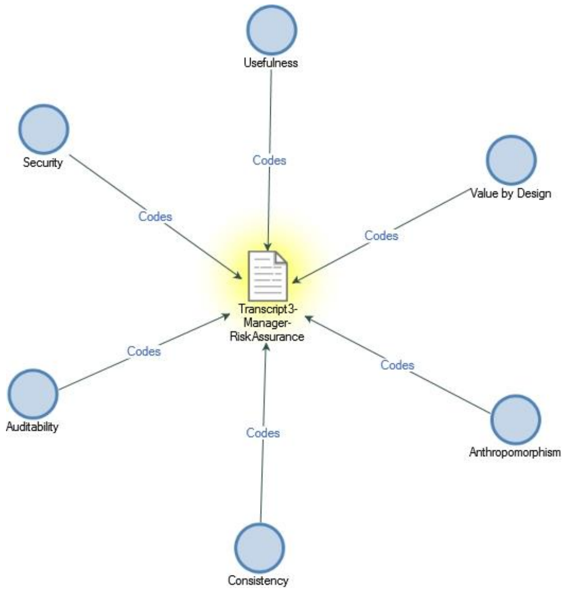
Auditability *goes together with* ***transparency*** *because if you have technology transparent, then it would be auditable.*

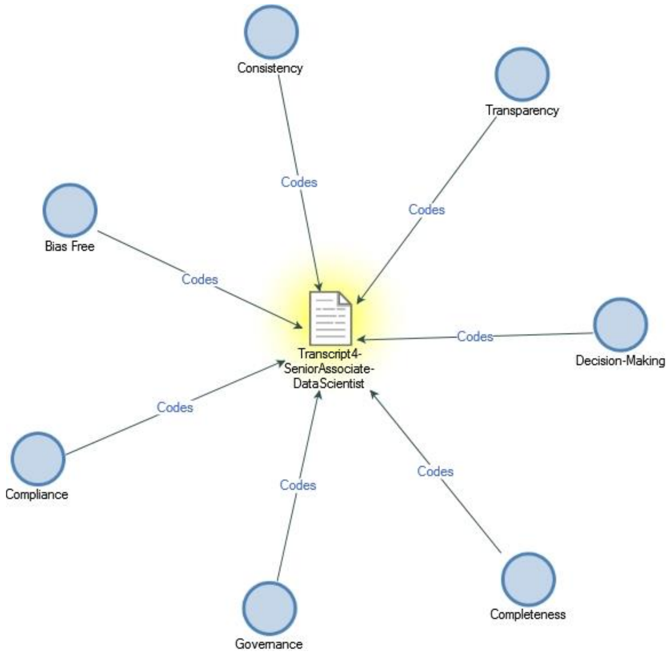
The perspective of a risk auditor on trust factors towards AI – Respondent 9

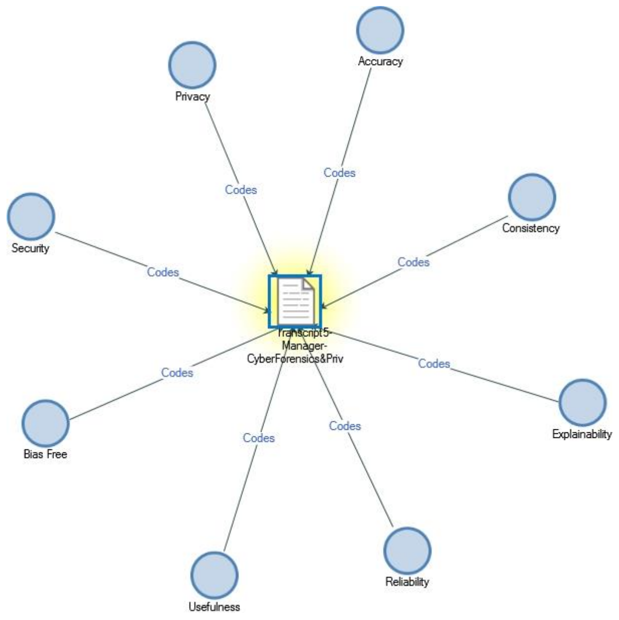
APPENDIX B - IDENTIFICATION OF CODES – INTERVIEW ANALYSIS

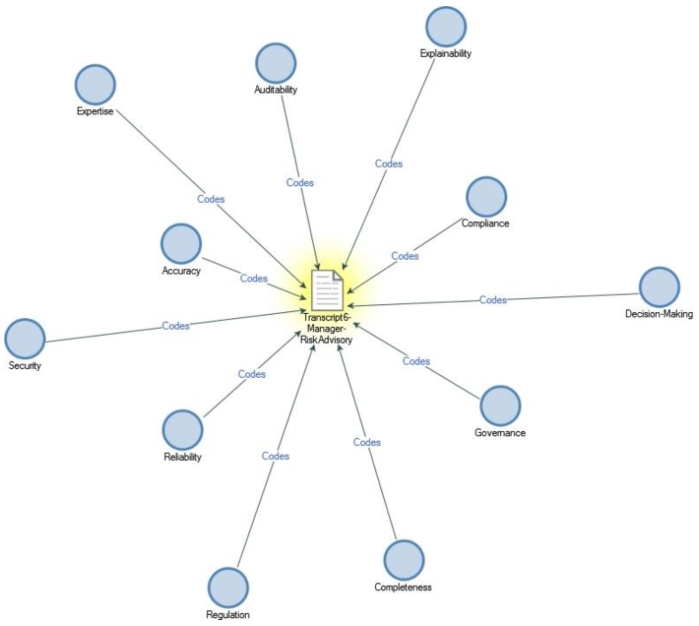


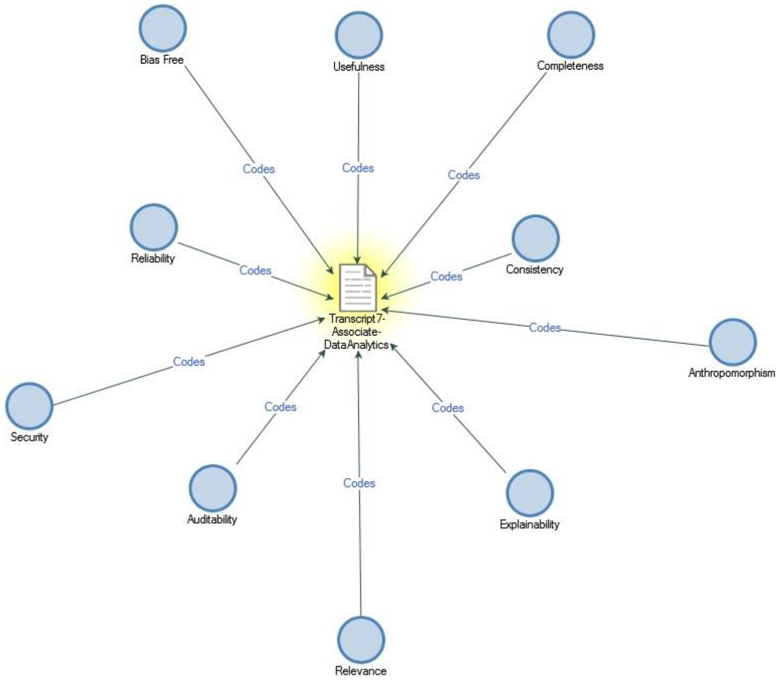


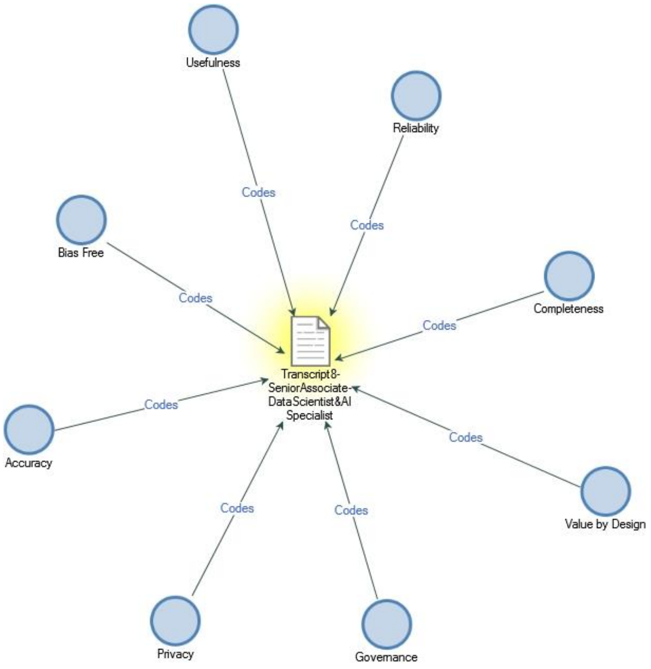


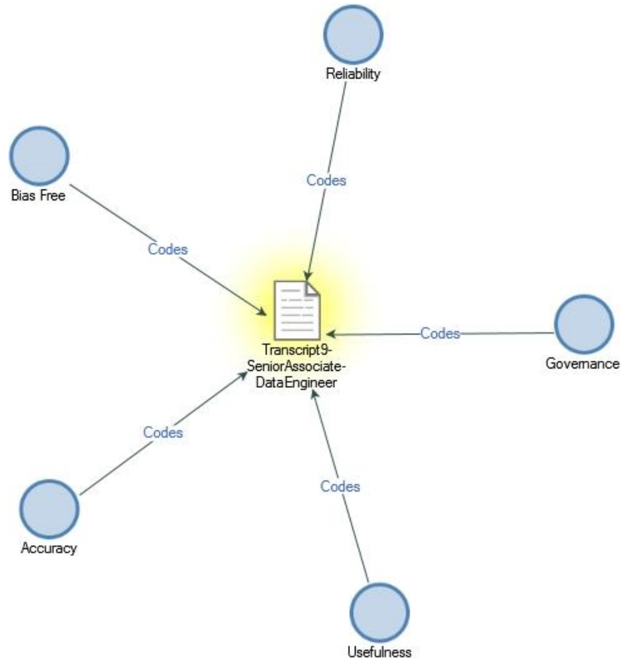


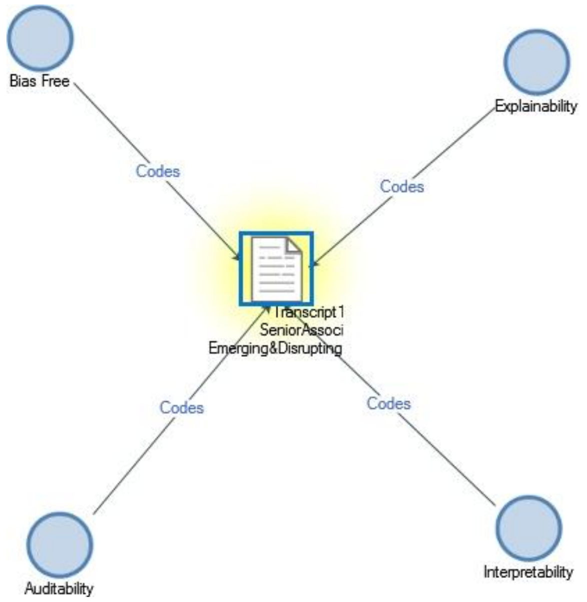

Transcript5-Manager-CyberForensics&Priv
Accuracy
Privacy
Security
Consistency
Bias Free
Explainability
Usefulness
Reliability
Codes

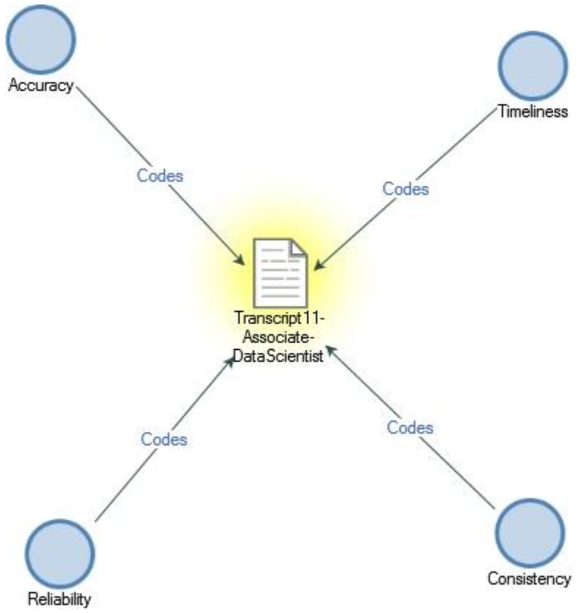

Transcript6-Manager-RiskAdvisory
Explainability
Auditability
Expertise
Compliance
Accuracy
Decision-Making
Security
Reliability
Governance
Completeness
Regulation
Codes

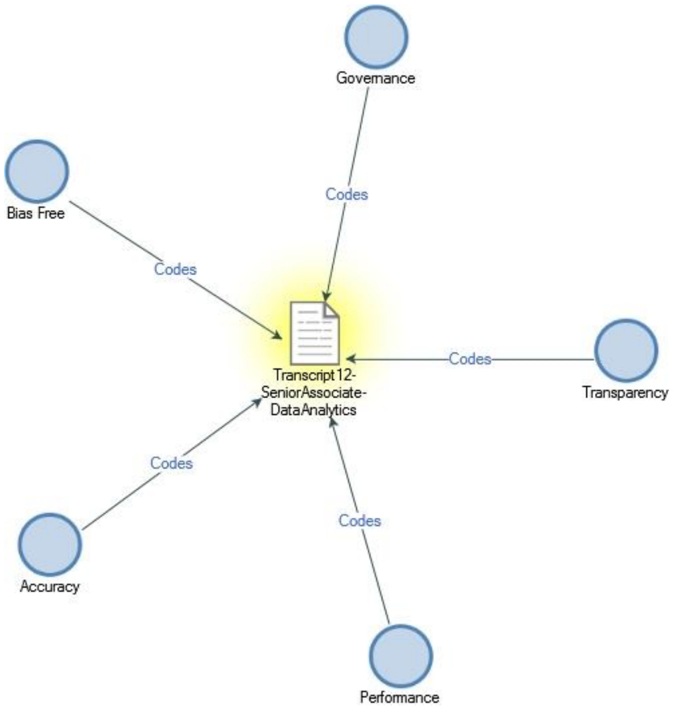

Transcript7-Associate-DataAnalytics
Bias Free
Usefulness
Completeness
Reliability
Consistency
Anthropomorphism
Security
Auditability
Explainability
Relevance
Codes

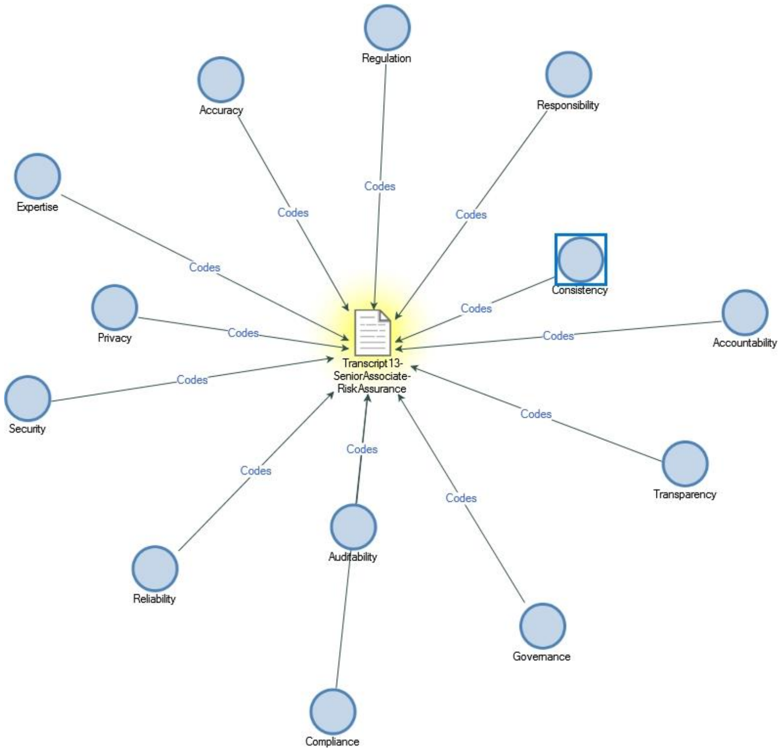

Transcript8-SeniorAssociate-DataScientist&AI Specialist
Usefulness
Reliability
Bias Free
Completeness
Accuracy
Value by Design
Privacy
Governance
Codes

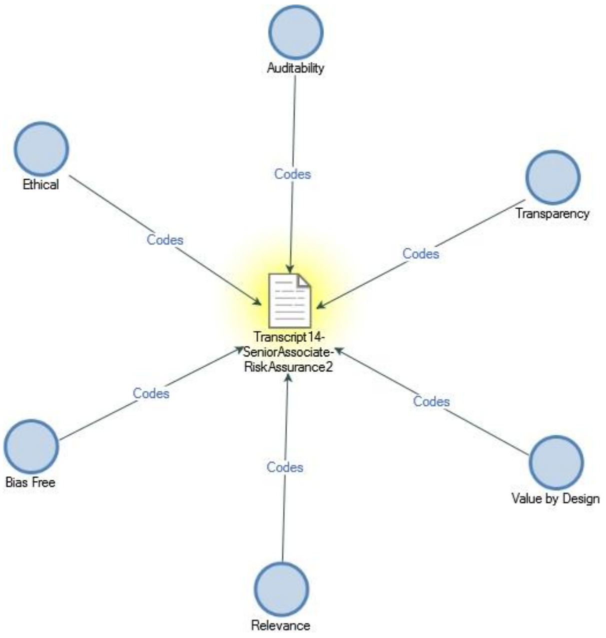

Reliability
Codes
Bias Free
Codes
Codes
Governance
Transcript9-
SeniorAssociate-
DataEngineer
Codes
Accuracy
Codes
Usefulness

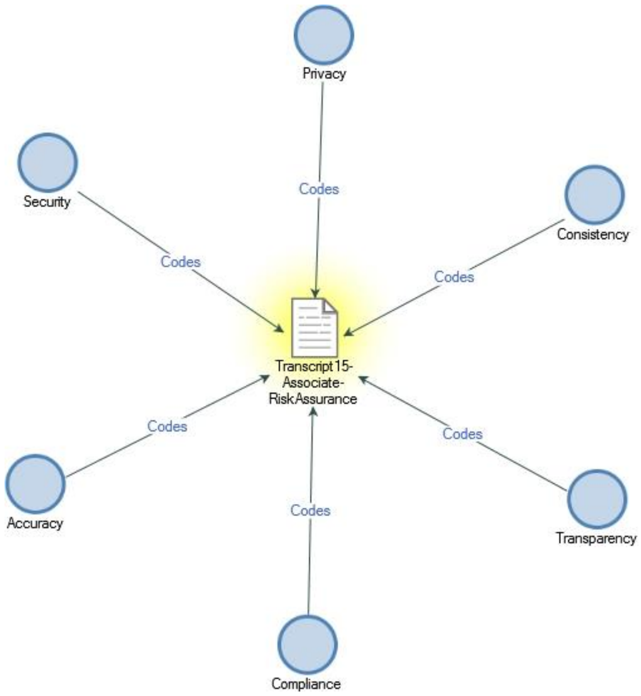

Bias Free
Explainability
Codes
Codes
Transcript1
SeniorAssoci
Emerging&Disrupting
Codes
Codes
Auditability
Interpretability

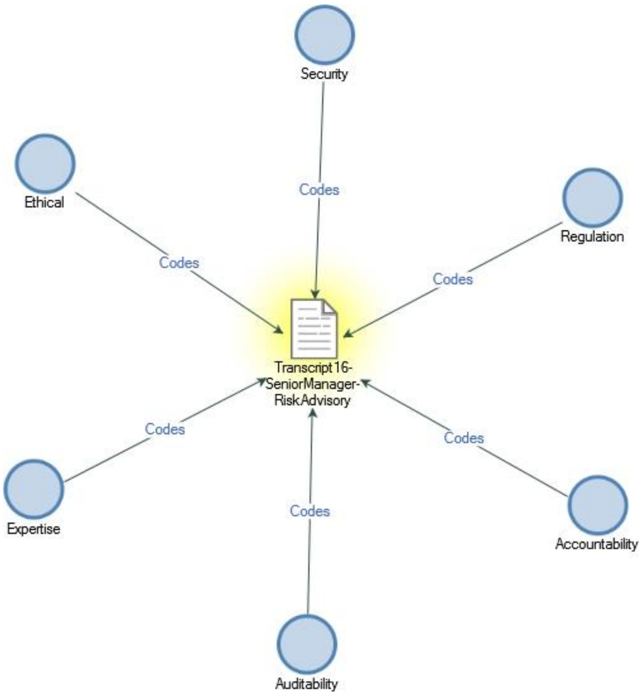

Accuracy
Timeliness
Codes
Codes
Transcript11-
Associate-
DataScientist
Codes
Codes
Reliability
Consistency


Governance
Codes
Bias Free
Codes
Transcript12-
SeniorAssociate-
DataAnalytics
Codes
Transparency
Codes
Accuracy
Codes
Performance

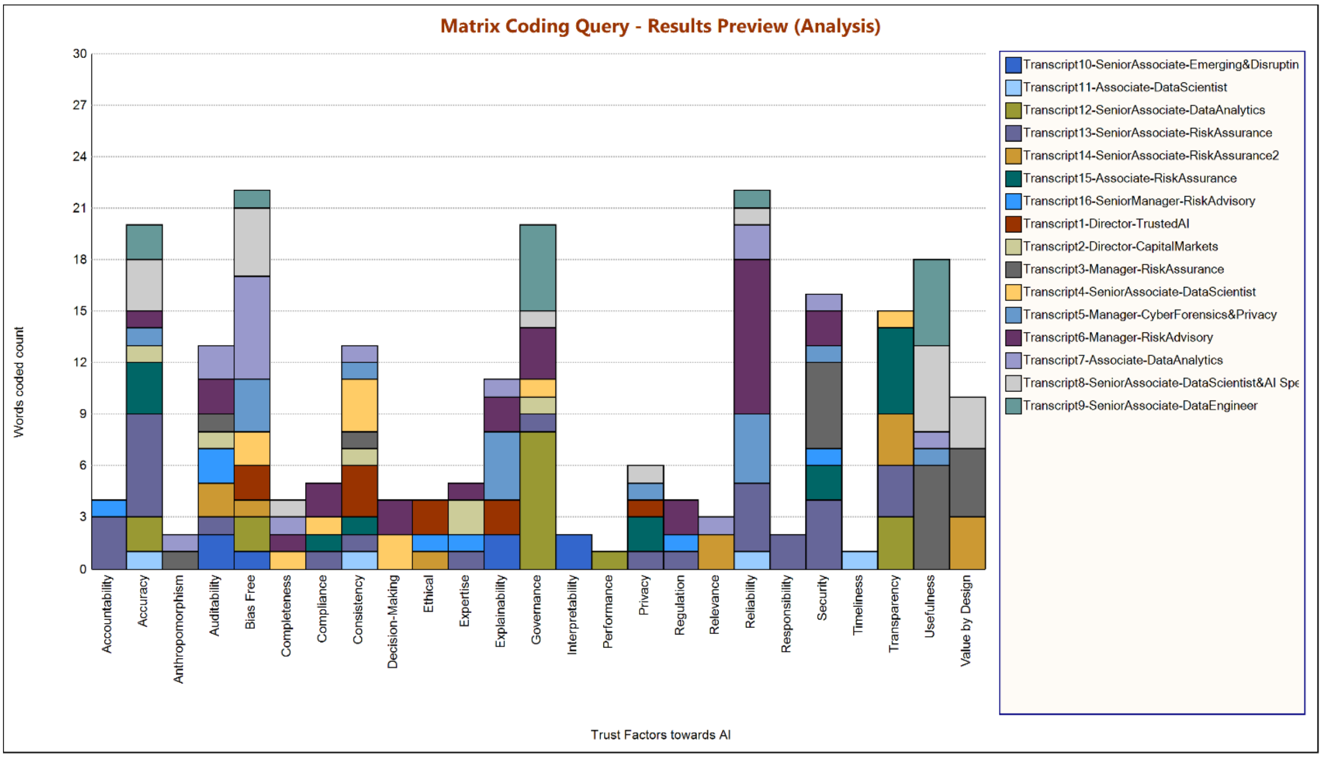

Regulation
Accuracy
Responsibility
Expertise
Consistency
Privacy
Accountability
Security
Transparency
Reliability
Auditability
Governance
Compliance
Codes
Transcript13-SeniorAssociate-RiskAssurance

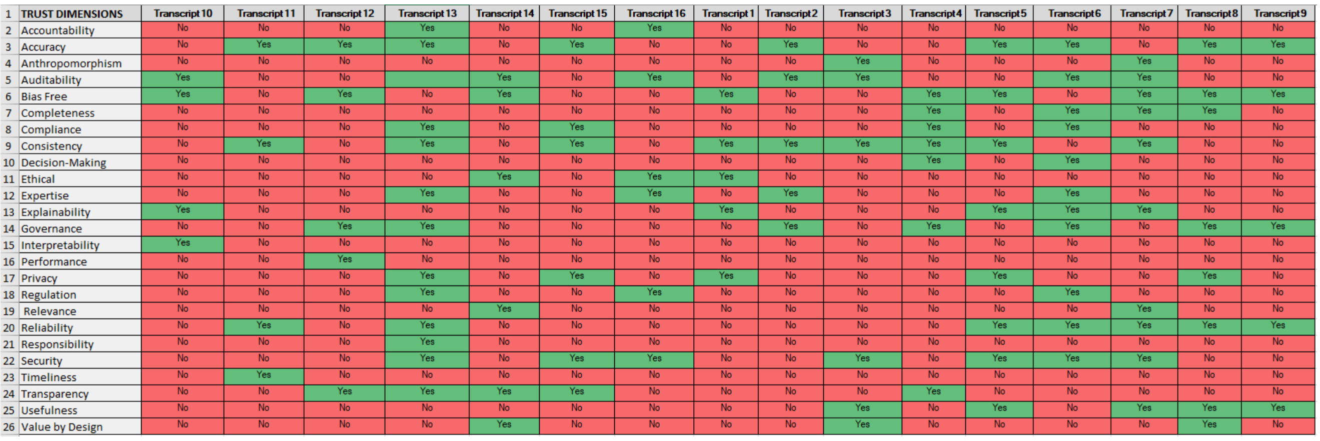

Auditability
Ethical
Transparency
Bias Free
Value by Design
Relevance
Codes
Transcript14-SeniorAssociate-RiskAssurance2


Privacy
Security
Consistency
Accuracy
Transparency
Compliance
Codes
Transcript15-Associate-RiskAssurance

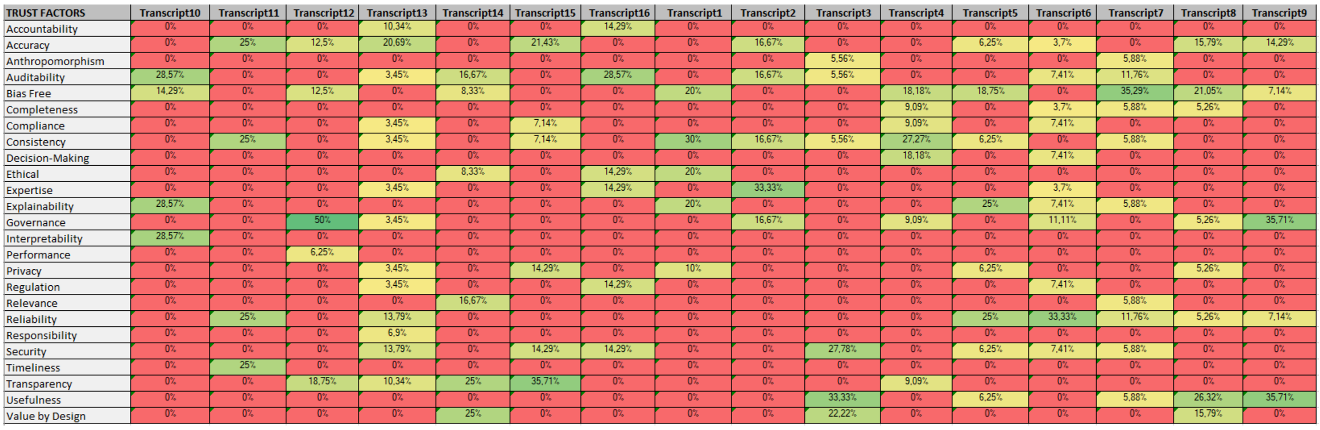

Security
Ethical
Regulation
Expertise
Accountability
Auditability
Codes
Transcript16-SeniorManager-RiskAdvisory

APPENDIX C - RESULTS OF THE CODED TRANSCRIPTS

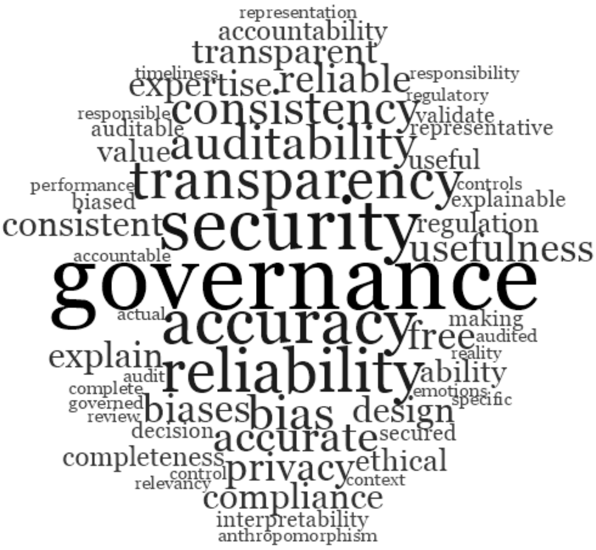




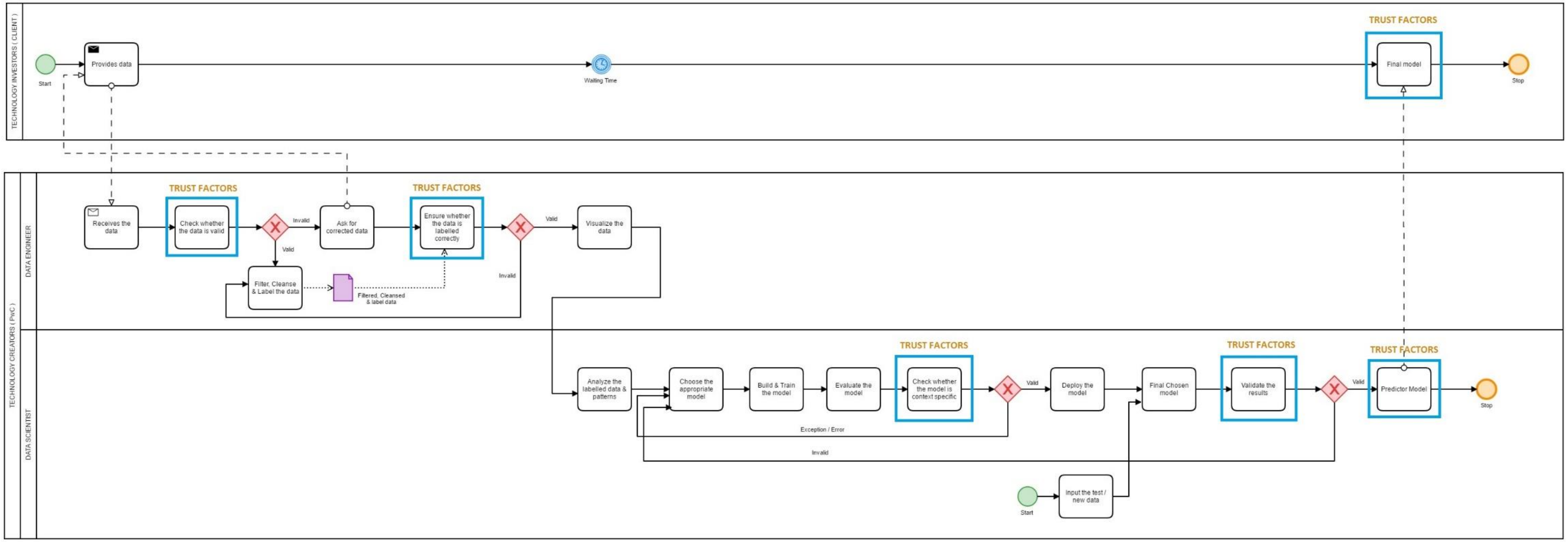
1	TRUST DIMENSIONS	Transcript 10	Transcript 11	Transcript 12	Transcript 13	Transcript 14	Transcript 15	Transcript 16	Transcript 1	Transcript 2	Transcript 3	Transcript 4	Transcript 5	Transcript 6	Transcript 7	Transcript 8	Transcript 9
2	Accountability	No	No	No	Yes	No	No	Yes	No	No	No	No	No	No	No	No	No
3	Accuracy	No	Yes	Yes	Yes	No	Yes	No	No	Yes	No	No	Yes	Yes	No	Yes	Yes
4	Anthropomorphism	No	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes	No	No
5	Auditability	Yes	No	No		Yes	No	Yes	No	Yes	Yes	No	No	Yes	Yes	No	No
6	Bias Free	Yes	No	Yes	No	Yes	No	No	Yes	No	No	Yes	Yes	No	Yes	Yes	Yes
7	Completeness	No	No	No	No	No	No	No	No	No	No	Yes	No	Yes	Yes	Yes	No
8	Compliance	No	No	No	Yes	No	Yes	No	No	No	No	Yes	No	Yes	No	No	No
9	Consistency	No	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	No
10	Decision-Making	No	No	No	No	No	No	No	No	No	No	Yes	No	Yes	No	No	No
11	Ethical	No	No	No	No	Yes	No	Yes	Yes	No	No	No	No	No	No	No	No
12	Expertise	No	No	No	Yes	No	No	Yes	No	Yes	No	No	No	Yes	No	No	No
13	Explainability	Yes	No	No	No	No	No	No	Yes	No	No	No	Yes	Yes	Yes	No	No
14	Governance	No	No	Yes	Yes	No	No	No	No	Yes	No	Yes	No	Yes	No	Yes	Yes
15	Interpretability	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No	No
16	Performance	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No
17	Privacy	No	No	No	Yes	No	Yes	No	Yes	No	No	No	Yes	No	No	Yes	No
18	Regulation	No	No	No	Yes	No	No	Yes	No	No	No	No	No	Yes	No	No	No
19	Relevance	No	No	No	No	Yes	No	No	No	No	No	No	No	No	Yes	No	No
20	Reliability	No	Yes	No	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes
21	Responsibility	No	No	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No
22	Security	No	No	No	Yes	No	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No
23	Timeliness	No	Yes	No	No	No	No	No	No	No	No	No	No	No	No	No	No
24	Transparency	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	No
25	Usefulness	No	No	No	No	No	No	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes
26	Value by Design	No	No	No	No	Yes	No	No	No	No	Yes	No	No	No	No	Yes	No



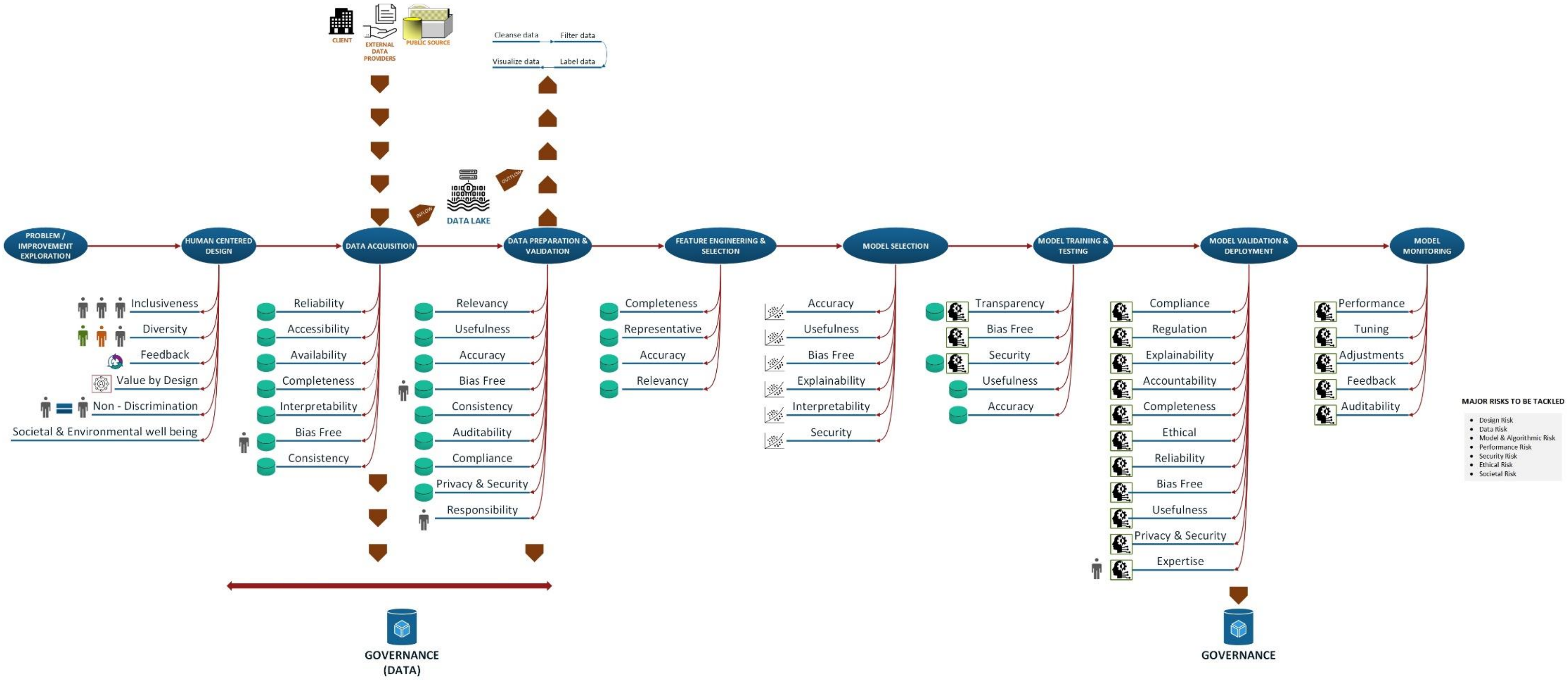
TRUST FACTORS	Transcript10	Transcript11	Transcript12	Transcript13	Transcript14	Transcript15	Transcript16	Transcript1	Transcript2	Transcript3	Transcript4	Transcript5	Transcript6	Transcript7	Transcript8	Transcript9
Accountability	0%	0%	0%	10,34%	0%	0%	14,29%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Accuracy	0%	25%	12,5%	20,69%	0%	21,43%	0%	0%	16,67%	0%	0%	6,25%	3,7%	0%	15,79%	14,29%
Anthropomorphism	0%	0%	0%	0%	0%	0%	0%	0%	0%	5,56%	0%	0%	0%	5,88%	0%	0%
Auditability	28,57%	0%	0%	3,45%	16,67%	0%	28,57%	0%	16,67%	5,56%	0%	0%	7,41%	11,76%	0%	0%
Bias Free	14,29%	0%	12,5%	0%	8,33%	0%	0%	20%	0%	0%	18,18%	18,75%	0%	35,29%	21,05%	7,14%
Completeness	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	9,09%	0%	3,7%	5,88%	5,26%	0%
Compliance	0%	0%	0%	3,45%	0%	7,14%	0%	0%	0%	0%	9,09%	0%	7,41%	0%	0%	0%
Consistency	0%	25%	0%	3,45%	0%	7,14%	0%	30%	16,67%	5,56%	27,27%	6,25%	0%	5,88%	0%	0%
Decision-Making	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	18,18%	0%	7,41%	0%	0%	0%
Ethical	0%	0%	0%	0%	8,33%	0%	14,29%	20%	0%	0%	0%	0%	0%	0%	0%	0%
Expertise	0%	0%	0%	3,45%	0%	0%	14,29%	0%	33,33%	0%	0%	0%	3,7%	0%	0%	0%
Explainability	28,57%	0%	0%	0%	0%	0%	0%	20%	0%	0%	0%	25%	7,41%	5,88%	0%	0%
Governance	0%	0%	50%	3,45%	0%	0%	0%	0%	16,67%	0%	9,09%	0%	11,11%	0%	5,26%	35,71%
Interpretability	28,57%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Performance	0%	0%	6,25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Privacy	0%	0%	0%	3,45%	0%	14,29%	0%	10%	0%	0%	0%	6,25%	0%	0%	5,26%	0%
Regulation	0%	0%	0%	3,45%	0%	0%	14,29%	0%	0%	0%	0%	0%	7,41%	0%	0%	0%
Relevance	0%	0%	0%	0%	16,67%	0%	0%	0%	0%	0%	0%	0%	0%	5,88%	0%	0%
Reliability	0%	25%	0%	13,79%	0%	0%	0%	0%	0%	0%	0%	25%	33,33%	11,76%	5,26%	7,14%
Responsibility	0%	0%	0%	6,9%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Security	0%	0%	0%	13,79%	0%	14,29%	14,29%	0%	0%	27,78%	0%	6,25%	7,41%	5,88%	0%	0%
Timeliness	0%	25%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Transparency	0%	0%	18,75%	10,34%	25%	35,71%	0%	0%	0%	0%	9,09%	0%	0%	0%	0%	0%
Usefulness	0%	0%	0%	0%	0%	0%	0%	0%	0%	33,33%	0%	6,25%	0%	5,88%	26,32%	35,71%
Value by Design	0%	0%	0%	0%	25%	0%	0%	0%	0%	22,22%	0%	0%	0%	0%	15,79%	0%



APPENDIX D - HIGH-LEVEL PROCESS FLOW IN AI DEVELOPMENT



APPENDIX E - INITIAL VERSION OF TRUSTED AI MODEL



APPENDIX F - DETAILED INDICATORS OF TRUST FACTORS IN A TRUSTED AI MODEL

Human-centered design phase

FACTORS	INDICATORS
Inclusiveness	• The potential AI users, developers, representative of industry sectors who might be impacted by AI are included. • Encouraging the involvement of key actors like data scientist, risk advisors, policy regulators to understand the challenges, requirements in the early design phase. • Domain-specific functional specialists are included apart from the techies to make sure that the resulting model is used effectively in the specific domains. • Defining the target group and ensuring whether all the representative of those targeted group is involved during the design. • A diverse set of representatives are present during the design phase in order to encourage the diversity of opinions.
Value by design	• The proposed solution is actually giving value to the clients. • The solution proposed would be the best solution to the business needs and to the society (For example, AI chatbot.).
Societal and environmental well being	• Seeking guidance from the experts, the target group of the solution in order to solve the problem and promote "AI for good". • Emphasizing the importance of privacy, security, and safety with the relevant actors in the early design phase. • Collaboration with private and public sector institutions when it comes to AI's Societal important. • The proposed solution is made to ensure that it would benefit the human and not harm them in any instance.

Data acquisition phase

FACTORS	INDICATORS
Reliability	• The contents present in the data set is credible enough to process. • The data has been generated from a trusted source.
Accessibility	• The data is readily accessible or retrievable from the intended source. • A higher degree of openness in accessing the intended data.
Completeness	• The mandatory & main attributes in the data are not null. • The data has all the possible states relevant to the user population. • The data is completely representative of the real-world state.
Consistency	• The source data is consistent even after a certain point in time. • The physical instance of the data is in accordance with some standard format. • The value and entries in the data are the same in all the case.

Data preparation and validation phase

FACTORS	INDICATORS
Relevancy	• The processed data is relevant to the specific domain and purpose of interest in a given context. • Processed data contains the required variable in the right form and a representative of population interest.

Consistency	• The processed data and actual data source are consistent & representative, have no contradiction and are compatible with the previous data. • The attribute values in the data have no ambiguities after data processing.
Bias-Free	• The biased data identified during the data acquisition and processing phase has been removed. • Data contains all the possible representation of the subject, object and has no inclinations towards a specific object or a thing, i.e no traces of biases, prejudiced, and partial data were found. • Determining the cause if any variance is detected so as to avoid biases.
Accuracy	• The processed data hasn't lost the structure during the data processing stages. • The processed data is certified and free of error. • The processed data is accurate, objective & is corresponding to a real-world context.

Feature selection

FACTORS	INDICATORS
Relevancy	• The chosen features are applicable and addressing the actual context of the goal.
Completeness	• The chosen features are representative of the entire population.

Model selection

FACTORS	INDICATORS
Interpretability	• The result that a model can produce is interpretable with respect to the data.
Accuracy	• The chosen model can produce correct predictions, decisions, and recommendations results if trained properly.
Bias-Free	• The chosen model is not pre-trained to show any biases and can work well if trained properly.
Explainability	• The model can provide some explanation for every certain output.

Model training and testing

FACTORS	INDICATORS
Transparency	• The model can provide some explanation for every certain output.
Usefulness	• The model can provide some explanation for every certain output.
Bias-Free	• The potential skews found during the training has been addressed. • Any skews observed during the testing were identified and addressed. • The training and testing data are free of biases.
Accuracy	• Testing data & training data has samples that represent all the targeted subjects, things, objects, etc. • The model is able to produce desired results based on the training. • The model is able to produce the same set of desired results when tested with new data (test data).

Model validation and deployment

FACTORS	INDICATORS
Explainability	• The AI model is able to explain every decision, recommendation, the prediction made. • The decision made by the system can be understood and traced by human beings.
Reliability	• The AI models are performing as intended. • The AI models are working properly when a new range of inputs is being applied.
Usefulness	• The resulting AI model had addressed the challenges and is providing value to the clients, society and the stakeholders involved.

Model monitoring

FACTORS	INDICATORS
Performance	• Identifying and addressing the potential weakness can improve the performance of the model. • Tuning and adjustments are made to improve the performance of the model.

APPENDIX G - DETAILED INDICATORS OF TRUST FACTORS AROUND THE DEVELOPMENT OF AI

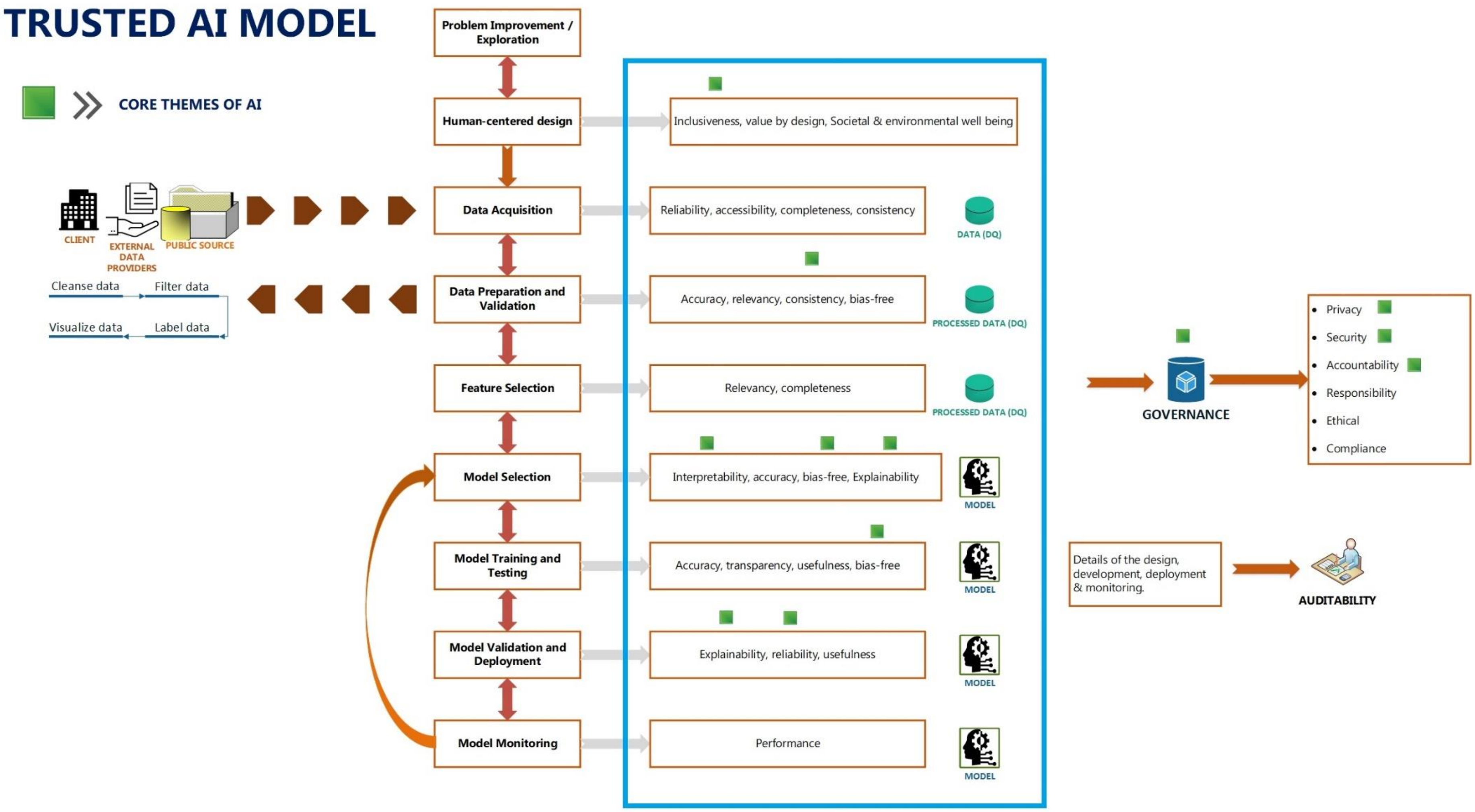
FACTORS	INDICATORS
Auditability	• Every change made with respect to datasets that have been documented is verified. • A process on how data was received, analyzed, cleansed, filtered, and labeled are documented and verified. • Assumptions made on the data to improve the usefulness of the data has been documented and reviewed. • Possible risks and opportunities that were encountered and how they were addressed is documented. • The results produced by AI systems are compared with the actual results. • Reviewing the output of the AI model and the meaning derived from those outputs. • Verify by interviewing the ones responsible for developing AI models that they are able to understand and explain the AI black box data. • Assessing the already existed data from the client or other third-party vendors to validate the usefulness of the data for requirements.
Governance	• A clear line of accountability, responsibility standards is clearly established. • Know the responsibility of the data & model owners at each and every phase involved in the creation of AI. • Ensuring that human is involved in every phase of the model as well as to oversee the overall activity of the AI systems. • Data protocols outlining 1. Who can access the data and under which circumstances. 2. How the data are being handled and how is it being protected. 3. Whether the data is in compliance with GDPR & data protection policies. 4. What kind of biases were identified, how were they solved? 5. Were there any measures been taken to inform the relevant stakeholders when major bases were notified? etc. should be put in place.

Under the governance, special attention has to be paid to factors like privacy and security

INDICATORS

- Identify and protect the core strategic data assets.
- Access to final processed data is restricted & secured from unauthorized access.
- Processed data has been encrypted & anonymized if any personal or confidential information is present.
- Processed data is used fairly and respects the privacy of the user's data.
- The processed data is in compliance with GDPR and other data protection policies.
- The chosen model can be resistant to malicious training.
- The model is making fair predictions with the data being trained
- The model is neither using the test data maliciously nor being fed with malicious data and is making fair predictions with the test data.
- The data used by AI systems or model is not used unlawfully or unfairly against the users.
- The AI systems are safe and secured and are not vulnerable to tampering or compromising the data that they have been trained on.
- AI models are protected from being exposed to unexpected situations.
- The AI models developed are limited to the context of the goal.
- The AI model respects the user privacy of information

APPENDIX H - IDENTIFICATION OF AI THEMES IN THE TRUSTED AI MODEL



TRUSTED AI MODEL

🟩 » CORE THEMES OF AI

Problem Improvement / Exploration
Human-centered design
🟩 Inclusiveness, value by design, Societal & environmental well being
CLIENT
EXTERNAL DATA PROVIDERS
PUBLIC SOURCE
Data Acquisition
Reliability, accessibility, completeness, consistency
DATA (DQ)
Cleanse data
Filter data
Visualize data
Label data
Data Preparation and Validation
🟩 Accuracy, relevancy, consistency, bias-free
PROCESSED DATA (DQ)
Feature Selection
Relevancy, completeness
PROCESSED DATA (DQ)
Model Selection
🟩 🟩 🟩 Interpretability, accuracy, bias-free, Explainability
MODEL
Model Training and Testing
🟩 Accuracy, transparency, usefulness, bias-free
MODEL
Model Validation and Deployment
🟩 🟩 Explainability, reliability, usefulness
MODEL
Model Monitoring
Performance
MODEL
🟩 GOVERNANCE
Privacy 🟩
Security 🟩
Accountability 🟩
Responsibility
Ethical
Compliance
Details of the design, development, deployment & monitoring.
AUDITABILITY