8-19-2022

Framing TRUST in Artificial Intelligence (AI) Ethics Communication: Analysis of AI Ethics Guiding Principles through the Lens of Framing Theory

Namrata Nagar
nn3631@rit.edu

Framing TRUST in Artificial Intelligence (AI) Ethics Communication:

Analysis of AI Ethics Guiding Principles through the Lens of Framing Theory

Namrata Nagar

School of Communication

College of Liberal Arts, Rochester Institute of Technology

A Thesis presented

in partial fulfillment of the

Master of Science Degree in Communication

Degree Awarded:

August 19, 2022

The members of the Committee approve the thesis of

Namrata Nagar presented on June 2, 2022.

Tracy Worrell, PhD
Professor and Director
School of Communication
Thesis Advisor

Cecilia O. Alm, PhD
Associate Professor
Department of Psychology
Committee Member

Eun Sook Kwon, PhD
Associate Professor and
Interim Director of Graduate Program
School of Communication

Tracy Worrell, PhD
Professor and Director
School of Communication

Table of Contents

Abstract

With the fast proliferation of Artificial Intelligence (AI) technologies in our society, several corporations, governments, research institutions, and NGOs have produced and published AI ethics guiding documents. These include principles, guidelines, frameworks, assessment lists, training modules, blogs, and principle-to-practice strategies. The priorities, focus, and articulation of these innumerable documents vary to different extents. Though they all aim and claim to ensure AI usage for the *common good*, the actual AI system outcomes in various social applications have invigorated ethical dilemmas and scholarly debates. This study presents the analysis of AI ethics principles and guidelines text published by three pioneers from three different sectors - Microsoft Corporation, National Institute of Standards and Technology (NIST), AI HLEG set up by the European Commission through the lens of media and communication's *Framing Theory*. The *TRUST Framings* extracted from recent academic AI literature are used as standard construct to study the ethics framings in the selected text. The institutional framing of AI principles and guidelines shapes the AI ethics of an institution in a soft (as there is no legal binding) but strong (incorporating their respective position/societal role's priorities) way. The AI principles' framing approach directly relates to the AI actor's ethics that enjoins risk mitigation and problem resolution associated with AI development and deployment cycle. Thus, it has become important to examine institutional AI ethics communication. This paper brings forth a *Comm-Tech perspective* around the ethics of evolving technologies known under the umbrella term - Artificial Intelligence and the human moralities governing them.

Keywords: Artificial Intelligence, Ethics, AI Principles, Framing Theory, AI TRUST Framings

Framing TRUST in Artificial Intelligence (AI) Ethics Communication

'Artificial Intelligence'(AI) can be understood as the simulation of human intelligence processes by computer systems. In addition, Rich and Knight (1991) in their book on AI defined it as "the study of how to make computers do things which, at the moment, people do better" (p.3). As AI techniques for handling large amounts of data developed and technology research progressed, AI capabilities to handle more and additional or new tasks increased (Rich et al., 1991). By the end of the last century, AI techniques were applied to large practical projects with huge data availability. Today, AI is proliferating fast and leading us all to the dawn of the fourth industrial revolution (Park, 2018). A wide range of social sectors are experiencing disruption due to newly developed and deployed AI applications. Key industries like healthcare, chemicals, aerospace, defense, agriculture, automotive, banking, insurance, media, entertainment, telecommunications, transport, retail, and travel (Chui et al., 2018) that mainly cater to societal needs are getting more involved in the AI race (Soni et al., 2020) for attaining better efficiency and productivity rates with the reduction in operational costs (Lewerenz, 2021).

Advance Artificial Intelligence Technology and the Black Box Problem

Right from Covid19 vaccine roll-out (AITrends, 2020) to self-driving vehicle technology (Grigorescu et al., 2019), medicine and nursing support in healthcare (Saniotis et al., 2020) to mainstream digital financial market (Mhlanga, 2020), movie recommendations on Netflix to shopping suggestions on Amazon, from robotic missions to space (Chien et al., 2006) to autonomous drones for fast climate change crisis assessment (Hernández et al., 2021), AI technologies are rapidly becoming the oil of our societal machinery. Deployed in a specific

sector (recruitment, medical, translation, etc.) as an application, a well-trained AI system can recognize patterns and make subsequent judgments or decisions with little to no human intervention (Broussard et al. 2019). Basically, AI is an umbrella term used to refer to computing systems that interpret input data, usually learn from them, use those learnings to complete specific tasks, and include new AI approaches like machine learning and deep learning (Benefo et al., 2022). However, AI technologies come with their own set of risks, complexities, and challenges. One of the many challenges to AI application developers is to minimize bias in the datasets that are used to train the AI models for specific tasks in the application area.

Known for handling huge volumes of data (big data) efficiently, AI systems are also known for the *bias problem* they inherit from their learning environment. Perez (2019) in her book 'Invisible Women' highlighted that there has been a gap in the *big data* about women historically. She argues that such data sets with historical gaps due to deliberate omissions of women's achievements, experiences, needs, and daily lives if used for advanced AI systems' training, would render a considerable population of the world (women) invisible. The author refers to it as 'brilliance bias'.

Several field experts have published their work regarding gender or other biases in AI training data for social applications. Obermeyer et al. (2019) demonstrated the presence of *racial bias* in a commercially used AI-based health system which predicted a reduced number of black patients for extra care compared to white patients having the same level of health risks. The researchers reported proprietary issues as a major hindrance in investigating the *how* and *why* of data bias in their study. As known gender disparities or a preponderance of men is

prevalent in many fields including research (Helmer et al., 2017). Caplar et al. (2016) quantified the prevalent *gender bias* (papers with women first-authors have citation rates pushed down by 10%) in the citations of astronomical publications using an advanced AI technique. Trained with historical research data (200,000 papers in 5 journals from 1950 to 2015), the AI model in the study *predicted* that the papers authored by women should have received 4% more citations than did those authored by men (Vesper, 2016). Such precisely stated gender disparity finding is unsettling but the *trust factor* on the AI model that is predicting this is more unsettling.

Similar research in various other fields has led to growing *trust* concerns regarding the reproduction of myriad kinds of biases: racial, brilliance, gender, representation (Caliskan et al., 2017) via people or institutions building AI systems or through training datasets (with noise, bias or gaps) or by the complex nature of the advances in AI technology. Facebook's 2017 case where Facebook AI Research Lab (FAIR) found their AI chatbots deviating from their programmed script to create their own incomprehensible language without human input (Bradley, 2017) is an example of deviation from *transparency* and *trust* together. The stakeholders' concerns are further exacerbated by evolving advanced AI technologies like deep learning (DL), an algorithmic system of deep neural networks, which on the whole remain opaque to human comprehension (Eschenbach, 2021). Incidents and discourses in these directions pose serious questions with respect to the usage of advanced AI technologies in social or mass consumer markets. They not just exacerbate fear of personal and private human data (accessible to advanced AI-based products like voice-based assistants - Apple Siri, Google Home, etc.; digital media or social micro-blogging sites like Twitter, Instagram, etc.) misuse but also strengthen the 'black-box' perception of AI. The concerns around AI or advanced AI technologies also touch upon issues of control, autonomy, intentionality, and responsibility.

Cases like *project maven* and ground applications like LAWS, AI-robots used in the military known as Lethal Autonomous Weapons Systems (Daisuke, 2019), point in these directions.

Ethics Principles, Guidelines and Leading Institutions' Core Approach towards Black Box AI:

AI applications indeed offer myriad opportunities for economic efficiency and quality of life, but the new forms of risks (Taeihagh, 2021) they generate need proper regulatory addressal by appropriate institutions and governing bodies. Well-formed, well-grounded, actionable ethics principles and guidelines vetted not just by developing institutions but government, academicians, and (concerned) public as well would be a good starting point in this direction.

This paper focuses on the framing of institutional ethics principles for artificial intelligence technology by three diverse pioneer institutions. Focusing on the *framing* of ethical principles and guidelines will bring out insights for governing the scale and speed of AI's socio-technical progress in society. So far, the big technology companies, some of them bigger than many countries' economies, are leading the AI-powered new industrial revolution across the globe. Their political role, control over AI development, and economic hold are notable at the national and international levels (Parviala, 2018). The ethics approach, manifested and communicated through AI principles of institutions like Microsoft, National Institute of Standards and Technology (NIST) and European Commission's High Level AI Expert Group (AI-HLEG) reflect the focus and action priorities with regards to new, emerging, transformational advance AI technologies like deep learning.

The three commanding institutions from three different sectors that are considered as

sample institutions for this study are-

Microsoft Corporation, a global organization with many artificial intelligence technology patents

The Office of the Chief Economist in Intellectual Property (IP) Data Highlights Report (2020) evaluated the AI diffusion with US patents from 1976 till 2018, coming to the conclusion that Big Tech companies and their AI technologies have made significant progress over the years. One of the pioneers in AI technology with various AI products and services backed by advanced research, Microsoft, owned 18,365 AI patents followed by IBM (with 15,046 AI patents), Samsung (11,243), Qualcomm (10,178), and Google (9536) in 2019 (iplytics, 2019). To avoid exposing their organization to serious financial, reputational, and legal risks, these tech giants have developed self-governing frameworks like *Responsible AI* by Microsoft, *AI Ethics* by IBM, *Google AI Principles* by Google and related oversight boards. Google published its AI principles in June 2018 as a charter to guide how they develop AI responsibly and the types of applications they pursue (Google, 2018) post discontinuing the controversial *Project Maven* contract with Department of Defense (9to5google.com, 2018). To follow-up and complement the internal governance structure and processes, the AI developing tech institution established an Advanced Technology External Advisory Council (ATEAC) with the goal to implement these AI principles. Alphabet's (Google parent) senior experts and board members added to their internal tech governance with a responsible innovation team that would consider and handle the most complex and difficult issues like 'AI Facial Recognition' (Google, 2019) which Google opted out of offering service until policy questions regarding *fairness* in machine learning (advanced AI technology) are settled. Such practices have become an important prerequisite for bidding on big contracts, especially involving government after

project maven where Google employees' protest eventually persuaded Google to end their involvement in an emerging military AI technology with a US Department of Defense initiative that seeked to leverage AI for automating drone footage analysis (Crofts & Rijswijk, 2020; Wilson 2020).

The National Institute of Standards and Technology (NIST), a non-regulatory federal agency within the U.S. Department of Commerce

Responding to the building pressure for AI standards and regulatory framework and considering the strategic importance of AI for nation's innovation, efficiency, equity, and security goals, the US federal government intervened in the AI market with federal policy (Presidential Memorandum, 2017; NAIIA, 2021; ASME, 2021). Following this federal policy initiative, the National Institute of Standards and Technology (NIST) published a notice and RFI in the Federal Register (84 FR 18490), about Artificial Intelligence Standards requesting public comments (NIST, 2019). NIST has now published Artificial Intelligence guidelines for US agencies and organizations that contribute to the nation's global leadership in the Artificial Intelligence technology space (Boulanin & Verbruggen, 2017; Susar & Aquaro, 2019).

Another forerunner in the global AI space, the European Union (EU) is one of the leading institutions in developing ethical guidelines for AI. The EU is active in the process of international norm establishment towards artificial intelligence technology, its purposes, and limits.

High-level Expert Group on Artificial Intelligence (AI HLEG), appointed by European Commission to act as the steering group of the European AI Alliance

Parviala (2018) studied the EU's artificial intelligence policy in detail from the lens of role theory. To understand the policy's internal as well as international implications and the EU's role in the global rise of AI, the researcher analyzed the European Commission's Coordinated Plan on AI and High-level Expert group on AI's draft Ethics Guidelines for Trustworthy AI. They arrived at the conclusion that the EU aspires to become a normative power in the new era of AI and is continuously working and gaining collaborations in the process of developing ethical guidelines for this transformative technology with a human-centric approach. The European Commission created an independent expert group in June 2018 to put forward guidelines for Trustworthy AI. Referred to as AI HLEG (High-Level Expert Group on AI), it presented the revised ethics guidelines after collecting feedback from open consultation during the piloting phase (2018-2019). These Ethics Guidelines for Trustworthy AI were made public by the European Commission on 8 April 2019 (EC, 2019).

Problem Statement

The discourse on AI ethics is dispersed over several themes as more and more institutions in the field release their ethics guidelines making the big-picture view of 'AI Ethics' landscape unclear and understanding regarding fair implementation difficult. Apart from the above selected pioneer institutions, several other user organizations, developer institutions, as well as government agencies have developed and published AI ethics principles and guidelines. Scholars are observing significant overlaps, differences, in these on-going attempts to create actionable ethical guidelines that mostly aim for 'the common good'. Currently, the AI ethics discourse is far from reaching widespread agreement on standard guidelines and regulatory

frameworks. The problem now lies in defining what is 'common good', 'social benefit' in our globalized and digitized world of today. Defining *AI for common good* comes to precisely defining not just fairness, human rights, widely honored values (eg: fairness, privacy) but also defining 'harm', AI use cases undermining human rights and values that are prioritized/ignored in social scenarios that relate directly to our economy and society.

Significance of this Study

Siau & Wang's (2020) categorization, and Hagendorff's (2020) institutional overlaps and omissions are issue-based evaluations of major AI guidelines and provide a semi-systematic overview of AI issues and normative stances in the evolving field of *AI ethics, governance, and regulation.* Scholars like Danaher (2018), Whittlestone et al. (2019) and Saetra et al. (2021) have been addressing ethical concerns, tensions in the ethics principles and need for actionable ethical AI principles in their research work.

Developing ethical principles, communicating them in the media, and analyzing them from different lenses can help bring to surface hidden tensions, new perspectives, tech-business-social priorities to the surface. This will help in improvisation, operationalization, and conflict resolution that happen as the AI tech and its social use-case scenarios evolve with time. The framing approach of any institution's AI principles enjoins risk mitigation and problem resolution associated with this emerging technology. This study brings together AI ethics and its communication which are rooted in the complete AI development and societal deployment or maintenance cycle. It contributes insights to the on-going AI ethics discourse for current and future AI societal applications requiring regulatory approach and assurance services ensuring

stakeholders' understanding regarding AI tech performance, risk, and compliance. An attempt has been made through this research study to establish what the roles of trust and understanding within the functions of advanced AI technologies and their associated mass (users/stakeholders/public) communication are and why they must be paid attention to. Here, media and communications *framing theory provides an apt methodology* to analyze institutional AI ethics principles and guidelines text framings.

Literature Review

Framing Theory Literature: Lens for the Study and Tool for AI Ethics Communication

Goffman (1986) was one of the first scholars to have developed the general concept of framing. He called frames the ‘‘schemata of interpretation,’’ a framework that helps in making an otherwise meaningless succession of events into something meaningful (p. 21). In simple terms, Goffman's work illuminates 'framing' as a mind tool through which people organize what they see in everyday life. Berger (1986) in his foreword for Goffman's seminal work on frame analysis (1974) wrote "There may, in short, be frames within frames within frames within frames" (p. 14). According to Entman (1993), to frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation" (p. 52).

Asplund (2014) in their research on climate change communication in the Swedish agricultural sector presented departure points regarding climate change understandings among Swedish public (especially farmers) perceptions and mainstream media representations. They

analyzed climate change frames and frame formation in Swedish agriculture magazines and conducted farmers' focus groups to find the contrasts. The study found Swedish farm magazines' framing climate change in terms of conflict, scientific uncertainty, and economic burden using metaphorical representations of war and games to form the overall frames of climate change. Whereas the farmers in the focus groups perceived climate change communication as an issue of credibility and thus their frames were about natural versus human-induced climate change supported with analogies, distinctions, keywords, metaphors, and prototypical examples based on both experienced and non-experienced arguments. In another qualitative study, Neill et al., (2017) utilized Entman's framing definition (1993) to study and measure 'attention' and 'prominence' of news items in print and TV to find the dominant frames in legacy and social media coverage of the IPCC Fifth Assessment Report (The Fifth Assessment Report of the United Nations Intergovernmental Panel on Climate Change). By qualitatively examining elite discourses, mass media research, and peoples' everyday perceptions of climate change frames, the researchers developed a frame schema inductively. IPCC data were examined for frames' constituent elements including metaphors, imagery, typical sources to identify and fully define frames available to journalists. This approach also helped in situating ten frames, identified and described in the study in a sociopolitical context.

Framing research that evolved from political science and sociology refers to the ''frames in communication'' (Chong & Druckman, 2007b, p. 106). Framing research on these foundations focuses on the ''words, images, phrases, and presentation styles'' (Druckman, 2001, p. 227) that are generally used to construct news stories and the processes that shape this construction. Chuan et al. (2019) analyzed the content of five major American newspapers (from 2009 to 2018) using framing theory to understand how artificial intelligence is being

framed in the U.S. print media. Statistically, the study identified three broad framings (AI Risks and Benefits Framing; Personal Vs. Societal Impact framing; Episodic vs. Thematic Framing) for artificial intelligence technology used by the US print media. In a quantitative study, Miller et al. (1998) performed a computer-assisted frame analysis of candidates in presidential primaries assuming frames are manifested in specific words. Their frame identification was guided by what they called 'frame-mapping' where they examined words that tend to occur together and identified frames with the help of clustering techniques. Thus, different scholars have shown various theoretical and operational understandings of frames in their work (Matthes, 2009). D'Angelo (2002) asserted that the way diverse theoretical and methodological approaches for framing contribute to the comprehensive understanding of *framing* and to the field of communication cannot be matched by a single framing paradigm.

Framing is related to the agenda-setting tradition but expands the research by focusing on the essence of the issues at hand rather than on a particular topic (Arowolo, 2017). The concept of framing holds similarities to concepts of the explanatory theme and discourse analysis (Neill et al., 2015). Here, it is important to understand that frames in communication reflect a speaker's (selected pioneer institutions in this study) emphasis while a frame in an individual's thought refers to what they believe to be the most salient aspect of the communication (Chong et al., 2007) be it an event, public speech, audio, video, or text. The societal understandings of the issue narrow around the most dominant or consistent frame/s in media (Foley et al. 2019). Scheufele (1999) identified four main processes in framing research (a) *frame building* (b) *frame setting* (c) *individual-level effect of frames* (d) *journalists as audience*. As it is not possible to incorporate too many different framing methodologies or theoretical approaches together, this study focuses on the first process which is *frame building*

grounding it in Entman's framing definition and following a methodical approach under 'Frames in Communication' (Chong and Druckman, 2007, p106-107) for identifying frames in AI ethics principles communication. Thus, *frame building* of an AI institution's ethics principles can be understood as the inclusion of n (1, 2, 3...) number of AI themes while excluding other prevalent themes. There can be sub-themes within dominant themes or other prevalent themes just mentioned as sideline text/topic in the AI ethics principles and guidelines text. It is a strategic communication choice made as per various interplaying factors in the market or society they operate in.

Academic Frames for this Study: TRUST Framings

The uncertain, unclear, and debilitating situation of getting unfair or biased or low-quality social outcome as a result of non-transparency or complex algorithmic functioning or flawed data usage in AI based social systems is commonly referred to as *the black box problem in AI* (Eschenbach, 2021; Innerrarity, 2021; Ratti and Graves, 2022; Carabantes, 2019, Michel 2020; Pipon et al., 2022). To address the black-box problems, field experts and developer institutions are exploring technical as well as ethics-based regulatory approaches to data processing, AI system learning, and social application development that can be used to minimize the risks and negative effects (like biased learning leading to unfair outcomes and noise or gaps in data leading to poor system performance). The *trust factor* on AI decision-making processes, predictions, and outcomes can be fully established only when the inner functioning of a model using techniques (such as machine learning, deep learning, and artificial neural networks (Jobin et al., 2019) for any social application is clear to stakeholders, fair to users, and transparent to actors in applied fields.

As mentioned in the problem statement of this study, several national, international, public, non-profit, academic institutions are developing and publishing ethics principles and guidelines to ensure AI is utilized for *common-good*. Private sector and public sector have developed specific offices and committees appointed or mandated to draft AI guiding documents to address the concerns and risks regarding AI. The US federal act - National Artificial Intelligence Initiative Act of 2020 (NAIIA) is one such example. It has led to formation of multiple AI dedicated offices or committees like National Artificial Intelligence Initiative Office, National Artificial Intelligence Advisory Committee, National AI Research Resource Task Force (ai.gov, 2022). Microsoft AI ethics efforts are led by its Aether Committee, the Office of Responsible AI (ORA) and Responsible AI Strategy in Engineering (RAISE) (Microsoft, 2022). Also, the appointment of a high-level expert group on AI (AI HLEG) by the European Commission are clear indications of how important the articulation of artificial intelligence principles and guidelines is.

The public communication of these AI principles and guidelines comes as the next important institutional step in the AI development to deployment cycle. Considering the transformative and proliferation power of AI, the starting point must be sound and good enough to support the end goal of delivering *common-good* in society. According to Goffman (1974) "...there will be a person-role formula. The nature of a particular frame will, of course, be linked to the nature of the person-role formula it sustains. One can never expect complete freedom between individual and role and never complete constraint" (p. 269). The person in Goffman's frame analysis essay can be a human or institutional actor that participates in an episode or series of communication framing activities. Goffman's person-role formula is applicable in AI ethics, regulation, and governance fields where diverse market players or institutional actors

develop their own set of AI guidelines and principles. Scholars are observing overlaps, differences, and tensions in these published AI ethics guidelines. Jobin et al., (2019) in their investigation regarding the constitution of *ethical AI* (such as ethical requirements, technical standards, and best practices required to realize Ethical AI) found that scholars diverge in interpreting AI principles, whom the principles pertain to, and their importance as well as implementation. Coming back to the problem statement of this study, the AI ethics discourse is right now far from reaching widespread agreement on standard guidelines and regulatory frameworks. Academic researchers' scholarly work can be one way of conceptualizing some important AI ethics framings whose inclusion in the AI ethics guiding documents would contribute to the development of ethical AI that works for *collective good*. The following **TRUST framings** are not exhaustive but provide crucial pointers on which the AI ethics principles and guidelines text of the selected AI players can be analyzed.

Transparent and Comprehensible AI Framing

To embed ethical principles into the design and deployment of AI-enabled systems, some form of *interpretability* in AI systems might be desirable or necessary. "*Transparency* and *explainability* of AI methods may therefore be only the first step in creating trustworthy systems" (The Royal Society, 2019). The Royal Society in its policy briefing (2019) clearly highlighted that there exists a range of approaches to *explainability* as there is a range of AI methods that are designed and deployed depending on the application at hand and hence there would be different explainability approaches serving different functions depending upon the AI application in focus. Intel researcher Wei Xu in his article (2019), while discussing a human-

centered AI perspective for human-computer interaction (HCI) professionals, reinforced the need for focusing on important non-technical factors like explainability and comprehensibility, apart from technical aspects in AI system development. The researcher proposed an extended HAI (Human-centered AI) framework to realize the goal of developing explainable and comprehensible AI. Realization of adequate human understanding by the AI system's capability to provide personalized explanations alongside the generated output is required for establishing trust (Greeff et al., 2021). Where, AI system's explanations relates to communicating the logic or intent behind its outcome to the user of the system. Thus, explainability of AI systems benefits identification of errors in output (if any) or isolation of undesirable outcomes due to gaps/issues in training data by the human users.

In this study, the AI principles and guidelines text that mentions any of the themes - Transparency, Explainability, Interpretability, Comprehensibility would be considered under *Transparent and Comprehensible AI Framing*.

Reliable and Safe AI Framing

Shneiderman (2020) in their discussion regarding Human-Centered Artificial Intelligence (HCAI), emphasized technical practices that support reliability and management strategies that create cultures of safety while developing high-performing trustworthy AI systems. According to Shneiderman "reliability is advanced by studying past performances by way of detailed audit trails, often called flight data recorders, which have been so effective in civil aviation. Ample testing and analyses of training data promotes reliable performance" (p. 496). For cultivating safety in the AI development process, the author suggested open

management strategies such as (a) leadership commitment to safety, (b) open reporting of failures and misses, (c) internal oversight boards for problems and future plans, and/or (d) public reports of problems and future plans.

In this study, the AI principles and guidelines text that mentions reliability, management practices directed towards safety, public reports of problems/failures/misses/future plans, or oversight boards would be considered under *Reliable and Safe AI Framing.*

User Control and Autonomy Framing

Shneiderman (2020) pointed to some key aspects relating to user control and autonomy regarding AI systems. First, discussing Sheridan and Verplank's (1978) one-dimensional list of 10 levels of control and autonomy (starting from complete human control to full computer autonomy) which they claimed to guide much of the tech research and development even today, the researcher proposed moving to Human-Centered Artificial Intelligence (HCAI) as an alternative. As per their analysis, this widely accepted levels of automation list only represents high automation or low human control situations. Thus, the researcher proposed to enable designers and developers to produce computer applications that would amplify, augment, enhance, and empower people while increasing automation via AI systems that are innovatively applied and creatively refined. Shneiderman stated Robin Murphy's law of autonomous robots as the law that captures the problem of autonomy which boils down to tension between autonomy and augmentation, asking for essential balance between user control and automation. Endsley (2018) gave guidelines for the design of human-autonomy systems that focused on human understanding of autonomous systems, minimizing complexity (usage of automated

assistance for routine tasks rather high-level cognitive function), and supporting situation awareness (like providing automation transparency with detailed explanation). Autonomy of AI based systems is often connected to *intelligence* of the machine in performing complex cognitive tasks like humans where humans have minimum to no control over decisions made by the AI system for the task (example: self-driving car control system where speed, motion, and direction are controlled by an AI system or autonomous military drones where unmanned aerial vehicle of any size operates without a pilot on board.)

Thus, AI principles and guidelines text mentioning autonomy, user/human control, human augmentation through automation, human consent are considered under *User Control and Autonomy framing.*

Secure and Privacy AI Framing

Just like ethics of collecting data for social science research involves data protection issues, ethics of data collection-to-use with artificial intelligence-based tech requires assurance of secure AI systems. Coeckelbergh (2020) in their book on AI Ethics wrote "An ethical use of AI requires that data are collected, processed, and shared in a way that respects the privacy of individuals and their right to know what happens to their data, to access their data, to object to the collection or processing of their data, and to know that their data are being collected and processed and (if applicable) that they are then subject to a decision made by an AI" (p.98).

Thus, AI principles and guidelines text mentioning any points regarding security and safety (w.r.t data collection, processing, access, share, consent, data subject to AI decision

making) embedded in AI systems that save users from hidden and effective form of manipulation, surveillance, and totalitarianism are considered under *Secure and Privacy AI Framing.*

The Other Framings

As more and more institutions, scholars, and government agencies join the AI ethics discourse and new AI approaches get developed for diverse social applications, new frames relating to AI complexity, concerns, and risks come to the surface. These evolving AI ethics, governance, and regulation discourses point towards new framings which require further in-depth study and exploration from different perspectives. Currently, they are included at shallower level during the AI principles and guidelines formulation compared to above explained framings. Thus, these relatively less emphasized framings are included under *The Other Framings.* Few prominently evolving framings are mentioned below-

Ethical Dilemma and Moral Framing. Siau and Wang (2020) questioned accountability, ethical standards, and software engineers' human rights laws knowledge while coding and developing AI systems in different categories. The article's category (B) talks about the ethics and morality of humans, which are themselves questionable (Bostrom and Yudkowsky, 2014). Thus, AI principles and guidelines text discussing moral behavior, ethical dilemmas or their resolution are considered *Ethical Dilemma and Moral Framing.*

Human Resource, Employment, Rights and Accessibility Framing. Siau and Wang's (2020) category (C) dealt with AI systems' consideration of democracy and civil rights, job replacement issues of human workers by automated systems, and accessibility of AI systems to the elderly and handicapped. On one side, advance AI technologies are shaping new human behaviors relating to easy to complex cognitive tasks while on other side these AI technologies are themselves evolving because of human (AI developers) choices, conflict resolution, and decisions. Thus, AI principles and guidelines text throwing light on relationship between AI technology and democracy or civil rights or future of human workforce are considered under *Human Resource, Employment, Rights and Accessibility Framing.*

Fairness, Non-discrimination, and Justice Framing; Inclusion, Diversity, Solidarity, Protection of Cultural Differences and Whistleblowers Framing. Hagendorff (2020) semi-systematically evaluated 22 of the major AI guidelines which according to their analysis were shaping the AI ethics discourse around that time. Classifying the AI guidelines under 22 key issues, they highlighted the overlaps and omissions in an institution's published AI guidelines with respect to other actors in the study. AI issue-based themes like privacy protection, transparency, openness, safety, cybersecurity, future of employment/worker rights, human autonomy, military, AI arms race, explainability, interpretability have been covered in the framing explanations above. Thus, principles texts mentioning fairness, non-discrimination, justice are considered under *Fairness, Non-discrimination, and Justice Framing.*

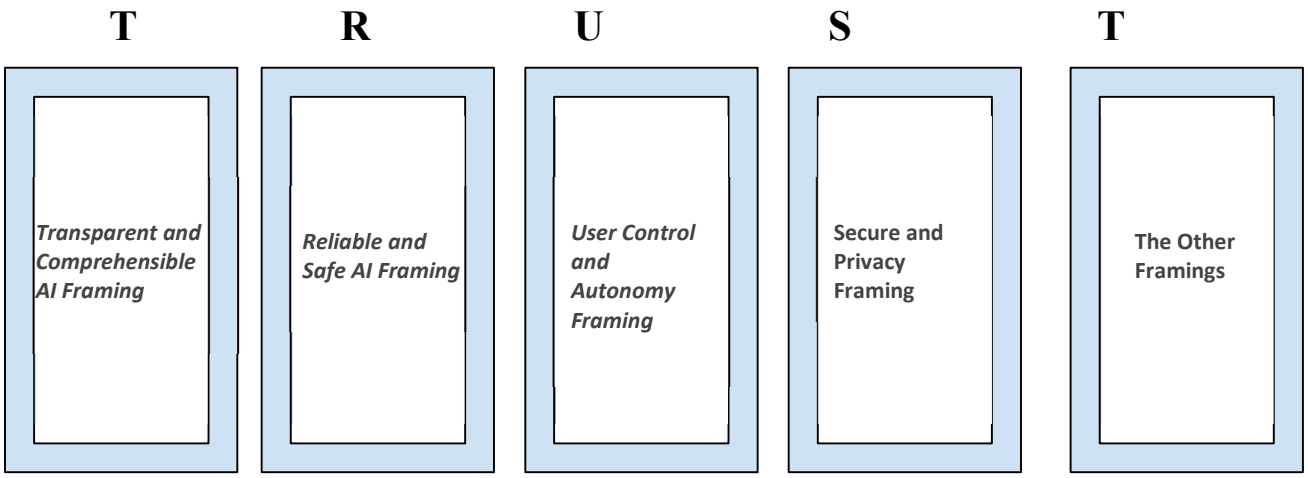
Accountability and AI Audits Framing. Several published AI guidelines and principles have references to responsibility, or 'Responsible AI' but according to Jobin et al.'s (2019) scoping review of then existing corpus of AI ethics guidelines and principles documents, responsibility and accountability are rarely defined. Thus, principles text explaining or elaborating on responsibility, accountability, or AI audits are considered under *Accountability and AI Audits Framing*.

There are AI ethics issues that both Siau and Hagendorff have touched upon. Additional issues/topics that Hagendorff (2020) brought forth are - responsible/intensified research funding, solidarity, inclusion, social cohesion, science-policy link, field-specific deliberations (health, military, mobility, etc.), diversity in the field of AI, certification for AI products, protection of whistleblowers, cultural differences in the ethically aligned design of AI systems, hidden costs (labeling, click work, content moderation, energy, resources), public awareness, education about AI and its risks. These are considered under *AI Education, Science Policy, and Public Awareness Framing;Responsible Research Funding, Hidden AI Costs, Field Specific Deliberations Framing.*

To sum up the above explained framings derived from literature review can be pictorially seen in Figure 1

Figure 1

*Academic TRUST Framings: **T**ransparent and Comprehensible AI Framing, **R**eliable and Safe AI Framing, **U**ser Control and Autonomy Framing, **S**ecure and Privacy AI Framing, **T**he Other Framings)*



Research Questions

The AI principles and guidelines framing analysis of the selected organizations is guided by the following research questions:

RQ1: What framings are observed in the selected institutions' AI principles and guidelines text?

RQ2: Which of the institutional framings are same or similar to TRUST (Figure 1) framings explained in this study? *(Where **TRUST** Framings indicate - **T**ransparent and Comprehensible AI Framing, **R**eliable and Safe AI Framing, **U**ser Control and Autonomy Framing, **S**ecure and Privacy AI Framing and **T**he Other Framings)*

Methodology

The purpose of this study is to analyze the AI principles and guidelines text of pioneer AI institutions (*Microsoft, NIST, AI-HLEG*) with diverse AI actor-role relationships (*Reference-*

Goffman's person-role formula) and find framings in their AI ethics communication. The framings are observed based on a set of framings (*TRUST Framings explained in previous section*) derived from the AI literature review. The selection of the sample AI organizations has been made to minimize personal bias by the author. Some other prominent AI organizations were excluded for examination either due to recent media coverages of ongoing ethics controversies (Example Google's Project Maven) or previous connections (with the author) or due to unclear institutional AI research approach which fuels AI innovation and also self-regulates or critically examines AI products and services for itself. The following section explains the text data collection and the researcher's analytic approach for frame identification in the selected institution's AI communication.

Phase 1

The researcher downloaded the text from each of the three institution's official website where their respective open access AI principles and guidelines were published. The source links to the studied text data is given in Table 1.

Data

AI Principles	Microsoft	AI-HLEG	NIST
Published Document Source Links	https://docs.microsoft.com/en-us/learn/modules/res	https://www.europarl.europa.eu/cmsdata/196377/AI%20HLEG	https://www.nist.gov/system/files/documents/2020/08/17/NIST

	ponsible-ai-principles/1-introduction	_Ethics%20Guidelines%20for%20Trustworthy%20AI.pdf	%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf
Active Web Links	https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6	https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence	https://www.nist.gov/artificial-intelligence
Document Length	13 full-length webpages with text	24 pages (August 2020) Draft NISTIR	36 pages (additionally 1 page

	on AI approach (7 video transcripts and 6 additional AI guideline blog entries) and 1 training module with 9 units	8312 and website updates on AI principles.	mentioning High Level Expert Group members) of Deliverable 1 (Ethics Guidelines for Trustworthy AI) and web links to Deliverables 2, 3, 4.

Table 1

Artificial Intelligence (AI) principles data for textual analysis as downloaded in Dec 2021

Phase 2

Matthes (2009) mentioned in their systematic analysis of media framing studies in the world's leading communication journals that frame analysis is an important methodology for examining the selection and salience of certain aspects of an issue. Drawing from Entman's (1993) framing definition and scholarly work mentioned in the literature review, framings were observed in the text-based data manually. A mix of inductive and deductive approach was adopted to identify the framings in the selected AI principles text. Rooted in qualitative paradigm of frame analysis, where frames are manifested in specific words, this study describes framings in depth with direct quotes from recently formulated and published AI principles and guidelines from selected AI pioneers that connects at different points or overlaps with on-going AI ethics scholarly discourse. The methodical steps given by Chong and Druckman (2007) in 'Frames in Communication' (p. 106) guide the frame identification during textual analysis of Microsoft.

NIST, and AI-HLEG's AI principles and guidelines' framings.

Before the analysis and findings are presented, it is necessary to specify how particular framing was identified. As per Chong and Druckman, "when researchers rely on computer programs to analyze large volumes of text, they must identify the universe of words that mark the presence of a frame" (p. 108). This study has identified the theme words (in the academic framing literature review section) that are indicative of the identified framings in the sample AI principles and guidelines text. The researchers acknowledged the identification of 'frames in communication' is about identifying the key considerations emphasized in a speech act. To do so, uniform measurement standards do not exist but according to Chong and Druckman, the most compelling communication studies follow four steps: (1) As a *frame in communication* can be defined only in relation to issues, or events or actors, identification of specific issue or event or actor is the first step in the process. (2) The second step involves isolation of specific attitudes, if the goal is to understand how *frames in communication* affect public opinion. (3) third step involves inductive identification of *initial set of frames for an issue* for creating a coding scheme. (4) Finally, in the fourth step the researcher selects the content sources for analysis based on the identified initial set of frames.

The researchers reviewed the meaning of the concept of framing, approaches to study framing and the effects of framing on public opinion. Considering the scope and goals of this study, apart from the second step which is to understand how frames in communication affect public opinion, all the above framing identifying steps are followed. The specific issues, supporting events, examples, AI actors, and selected sample institutions are identified and explained in the preceding sections. An initial set of framings corresponding to issues discussed

are identified and explained in the academic framing literature review section. Regarding Chong and Druckman's last step, the selection of AI principles and guidelines text from three institutional sources for analysis has been explained in the introduction section of this study.

The following section presents the analysis and findings based on the TRUST framings (Figure 1).

Analysis and Research Findings

As mentioned above, the framing approach of any institution's AI principles enjoins risk mitigation and problem resolution associated with this emerging technology. This understanding relates to Goffman's person-role formula where the nature of an AI actor's frame is directly linked to the role it sustains in society. The AI principles and guidelines framings shape the ethics of an institution in a soft (as there is no legal binding) but strong (incorporating their respective position/societal role's priorities) way. The AI ethics principles and guidelines text analysis is presented as answers to the two research questions asked in this study:

RQ1: What framings are observed in the selected institutions' AI principles and guidelines text?

With the aim to build trust in the AI system's entire life cycle (from development to deployment; from planning & communication to policy & investment recommendations), the High Level Expert Group on Artificial Intelligence (AI HLEG) appointed by European Commission developed a detailed guiding document that is now shaping Europe's overall AI approach to empower, benefit and protect European citizens (EU, 2022). Apart from the guidelines, termed as 'Ethics Guidelines for Trustworthy AI', the expert group provided three

more deliverables: Policy and Investment Recommendations for Trustworthy AI; Assessment List for Trustworthy AI (ALTAI); and Sectoral Considerations on the Policy and Investment Recommendations in the same AI Ethics Guidelines document. The AI ethics guidelines form the foundation over which other extended documents are built upon in detail. Every above stated extension is given a full chapter treatment after the *Ethics Guidelines foundation chapter*. The guidelines drafted by the AI high-level expert group is fundamentally grounded in Ethics in Science and New Technologies and the Fundamental Rights Agency with three necessary components - compliance with law, fulfillment of ethical principles and assurance of 'robustness' (from AI HLEG's EU documents and assessment list for trustworthy AI, it is specifically 'Technical Robustness' combined with AI system's safety, risk assessment to humans/animals/environment in various settings and fall back plans).

The guidelines identify key requirements which as Jobin et al. (2019) mentioned are stated to be non-binding. Though the seven requirements do not create any new legal obligations but provide detailed persuading advice to developers and stakeholders for adherence. The argument is that fulfilling AI HLEG's seven stated requirements would lead to development and deployment of AI systems that would be considered trustworthy. According to the guidelines the AI applications would be rendered trustworthy if they respect (1) Human agency and oversight (2) Technical robustness and safety, (3) Privacy and data governance, (4) Transparency, (5) Diversity, non-discrimination, and fairness, (6) Societal and environmental well-being, and (7) Accountability. The guidelines text and its communication to European parliament (EC, 2019) relates to the *Transparent and Comprehensible AI Framing*, *Reliable and Safe AI Framing*, *User Control and Autonomy Framing*, *Secure and Privacy AI* Framing, and *The Other Framings* (Diversity, Non-discrimination and Fairness, Accountability) of this study. Some example quotes

from selected AI principles and guidelines data documents mapped to TRUST framings of this study can be found in the Table 2 below. For more AI ethics text framings examples from EU's AI HLEG, Microsoft and NIST's AI principles and guidelines refer to Appendix A, Table 3 and Table 4 in the following sections.

Framing	**Identifying Word/Phrase**	**Examples**
***T**ransparent and Comprehensible AI Framing*	Transparency, Explainability, Interpretability, Comprehensibility	"Per-decision explanations provide a separate 370 explanation for each decision…Self-explainable models of machine learning systems themselves can be used as global explanations (since the models explain themselves). Likewise, many global explanations (including self-explainable models) can also be used to generate per-decision explanations." (NISTIR 8312, 2020, p.8)
***R**eliable and Safe AI Framing*	Reliability, Management Practices directed towards Safety, Public reports of Problems/Failures/Misses /Future plans, Oversight Boards	"ORA [Office of Responsible AI] puts Microsoft principles into practice by setting the company-wide rules for responsible AI through the implementation of our governance and public policy work. It has four key functions." "Aether [AI, Ethics and Effects in Engineering and

		Research] advises our leadership on the challenges and opportunities presented by AI innovations." "Responsible AI Strategy in Engineering (RAISE) is an initiative and engineering team built to enable the implementation of Microsoft responsible AI rules and processes across its engineering groups." (Microsoft Website: Microsoft on operationalizing Responsible AI)
***U**ser Control and Autonomy Framing*	Autonomy, User/Human Control, Human Augmentation through automation, Human Consent	"The fundamental rights upon which the EU [European Union] is founded are directed towards ensuring respect for the freedom and autonomy of human beings…AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills." (AI HLEG, 2019, p.12)
***S**ecure and Privacy AI Framing*	Security and Safety (w.r.t Data Collection, Processing, Access, Share, Consent, Data	"Prevention of harm to privacy also necessitates adequate data governance that covers the quality and integrity of the data used, its relevance in light of the domain in which the AI systems will be

	subject to AI Decision Making)	deployed, its access protocols and the capability to process data in a manner that protects privacy." (AI HLEG, 2019, p.17)
***T**he Other Framings*	Ethical Dilemma and Moral Framing, Fairness, Non-discrimination, and Justice Framing, Accountability and AI Audits Framing, AI Education, Science Policy, and Public Awareness Framing; Responsible Research Funding, Hidden AI Costs, Field Specific Deliberations Framing	"Beyond developing a set of rules, ensuring Trustworthy AI requires us to build and maintain an ethical culture and mind-set through public debate, education and practical learning." (AI HLEG, 2019, p.9) "The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension" (AI HLEG, 2019, p.9)

Table 2

Examples of Identified Framings in the Institutional AI Ethics Principles and Guidelines Text data (EU's AI HLEG, Microsoft, NIST)

Transparent and Comprehensible AI Framing

AI systems in social settings built on advanced AI techniques can be complex, thus NIST, a non-regulatory federal agency within the U.S. Department of Commerce whose mission is to

promote innovation and industrial competitiveness in the US, emphasizes 'transparency' in its AI principles. Three of four NIST AI principles are founded on transparency of AI system and their comprehensibility to human recipients of the information. Elaborating on types, meanings, and/or accuracy of explanations, NIST's AI principles validate the Royal Society's (2019) point regarding the existence of a range of explainability approaches (discussed under the *Transparent and Comprehensible AI Framing* in literature review). NIST's principles reiterate that depending on the application at hand and type of AI method designed and deployed in a social setting, the type and details of an explanation would vary. Microsoft's published case studies and video transcripts' text under AI principles cover *Transparent and Comprehensible AI Framing* (communicated with words -Transparency and Explainability), *Secure and Privacy AI Framing* (communicated with words- Privacy and Security) and *The Other Framings* (communicated with words - Fairness, Inclusiveness, Accountability) as discussed in the academic frames section of this study's literature review (for data examples refer Table 3 and Table 4).

Reliable and Safe AI Framing

An institutions' AI ethics guiding documents are non-legislative policy instruments or soft law, whose content/text is not legally binding but persuasive in nature (Jobin et al., 2019). Microsoft operationalizes its AI principles that it termed as 'Responsible AI' through its three offices/committees: the Office of Responsible AI (ORA), Aether Committee (Aether, which stands for AI, Ethics and Effects in Engineering and Research), and Responsible AI Strategy in Engineering (RAISE). RAISE is an initiative and engineering team built to enable the implementation of Microsoft responsible AI rules and processes across its engineering groups while Aether Committee makes recommendations on responsible AI issues, technologies, processes, and best practices to Microsoft's senior leadership (Microsoft, 2022). In short,

Microsoft applies their responsible AI principles with guidance from committees that advise its leadership, engineering, and every team across the company. Thus, its six core AI principles text first relate to Shneiderman's (2020) suggestion of open management strategies where leadership commits to safety and relies on internal oversight boards for problem resolution and planning for the future as discussed under *the Reliable and Safe AI Framing* in this study. On the other hand, NIST principles are designed to help a user/human (individuals, organizations, and society associated with AI) to understand AI systems, their explanation types, requirements, knowledge limits, outcome reliability, and prediction capabilities. According to NIST, Explainability as one property will contribute to characterizing trust in AI along with several other properties including Resiliency, Reliability, Bias, and Accountability. As AI system's explainability, resilience, reliability, bias, and accountability are communicated as properties, *the property of explainability* is considered a framing here as NIST's AI principles and guidelines text closely tie to this one property throughout while other properties are just mentioned in the document beginning. Though NIST has introduced a new principle named *The Knowledge Limits Principle* which precisely targets to achieve reliability in advance AI systems (Refer Table 3 and Table 4 for examples), this fourth principle deals with reliability frame at a shallow level.

User Control and Autonomy Framing

Focused on explanations regarding AI system's outcomes, behavior, and predictions provided to the end user/human recipients in a given situation for a task in hand, NIST's AI principles expand further to provide stakeholders an assessment framework for an AI solution's explanation and reliable outcome. The draft's fourth principle suggests defining the knowledge limits of AI-based machine systems that work as decision aids to humans clearly. According to

NIST putting this principle to practice would help in building trust among AI system users while minimizing cases of unjust AI decisions or outputs. This would also reduce AI dangerous or misleading outcome cases and ultimately leading to better and reliable outcomes. Explained with an example of bird classification AI system this NIST principle, points in the direction of human performance augmentation in tasks/projects through automation at much shallow level.

Secure and Privacy AI Framing and **The Other Framings.** Refer Table 3 and Table 4 for Microsoft and NIST's AI ethics principles and guidelines text examples (mapped to TRUST framings). For EU's AI-HLEG AI ethics principles and guidelines text examples (mapped to TRUST framings) refer to Appendix A.

Some New Framings

New framings found during the AI ethics and guidelines text analysis are *Microsoft's Non-Ableist Framing* that talks about designing, developing, and testing AI systems from a non-ableist perspective; *NIST's Knowledge Limits framing* that talks about preventing misleading, dangerous, or unjust AI decisions or outputs apart from building resilient and bias-free AI systems; *EU's AI HLEG's Societal and Environmental Well-being Framing* which talks about prevention of harm to beings, risk assessment on democracy, how AI affects human mind and the rule of law and distributive justice.

*RQ2: Which of the institutional framings are same or similar to TRUST (Figure 1) framings explained in this study? (Where **TRUST** Framings indicate - **T**ransparent and Comprehensible AI Framing, **R**eliable and Safe AI Framing, **U**ser Control and Autonomy*

*Framing, **S**ecure and Privacy AI Framing and **T**he Other Framings)*

A snapshot of the identified framings of the three institutions' AI ethics principles and guidelines with quoted examples from AI principles and guidelines data text is given in Table 3 and Table 4. These tables provide answer to research question 2. For EU's AI HLEG AI principles and guidelines text examples refer to Appendix A.

Microsoft	**NIST**	**AI-HLEG**
Transparent and Comprehensible AI Framing	Transparent and Comprehensible AI Framing (Explainability)	Transparent and Comprehensible AI Framing (Explicability)
Reliable and Safe AI Framing	Reliable and Safe AI Framing	Reliable and Safe AI Framing
NA	NA	User Control and Autonomy Framing
Secure and Privacy AI Framing	NA	Secure and Privacy AI Framing
The Other Framings (Fairness, Inclusiveness,	The Other Framings (Accountability)	The Other Framings (Fairness, Inclusiveness for

Accountability)		Vulnerable Groups or People at Risk of Exclusion or Historically Disadvantaged Groups, Accountability)
Not to take an Ableist Perspective (while designing, developing, or testing AI systems.)	Knowledge Limits Principle (Prevents Misleading, Dangerous, or Unjust Decisions or Outputs), Resiliency, Bias	Societal and Environmental Well-being, Prevention of Harm, Risk Assessment on Democracy, Human Mind, The Rule of Law and Distributive Justice

Table 3

Identified Framings in the Institutional AI Ethics Principles and Guidelines Text Data

TRUST Framings	**NIST**	**Microsoft**
Transparent and Comprehensible AI Framing	NA	NA
Transparency	NA	"At Microsoft, we've recognized six principles that we believe should guide AI

		development and use — fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability" (microsoft.com/en-us/ai/)
Explainability	"We introduce four principles for explainable artificial intelligence (AI) that comprise the fundamental properties for explainable AI systems. They were developed to encompass the multidisciplinary nature of explainable AI, including the fields of computer science, engineering, and psychology. Because one size fits all explanations do not exist, different users will require different types of explanations. We present five	NA

	categories of explanation and summarize theories of explainable AI." p.i	
Interpretability	NA	NA
Comprehensibility	NA	NA
Reliable and Safe AI Framing	NA	"It's important to recognize that as new intelligent technology emerges and proliferates throughout society, with its benefits will come unintended and unforeseen consequences, some with significant ethical ramifications and the potential to cause serious harm. It's our responsibility to make a concerted effort to anticipate and mitigate the unintended consequences of the technology we release into the world through deliberate

		planning and continual oversight. (Unit 3 of Identify guiding principles for responsible AI module) We are operationalizing responsible AI across Microsoft through a central effort led by the Aether, ORA, and RAISE. Together, Aether, ORA, and RAISE work closely with our teams to uphold Microsoft's responsible AI principles in their day-to-day work." (microsoft.com on operationalizing responsible AI across Microsoft)
Reliability	"This Knowledge Limits principle states that systems identify cases they were not designed or approved to	"To build trust, it's critical that AI systems operate reliably, safely, and consistently under normal

	operate, or their answers are not reliable." p.4	circumstances and in unexpected conditions." (Identify guiding principles for responsible AI module, Abstract)
Management practices directed towards Safety	NA	NA
Public Reports of Problems/Failure/Misses/Future Plans	NA	"...within 24 hours users realized that she [AI based Chatbot named Tay] could learn and began to feed her bigoted rhetoric, turning her from a polite bot into a vehicle for hate speech. This experience taught us that while technology may not be unethical on its own, people do not always have good intentions and we must consider the human element when designing AI systems. We learned to prepare for new

		types of attacks that influence learning datasets, especially for AI systems that have automatic learning capabilities." (Unit 3 of Identify guiding principles for responsible AI module, Section: Novel Attacks)
Oversight Boards	"The first appointments to the National Artificial Intelligence Advisory Committee (NAIAC) have been announced. The 27 experts will advise the President and the National AI Initiative Office on a range of issues related to artificial intelligence (AI). The committee will hold its first meeting on May 4, 2022, which will be open to the	NA

	public via webcast." nist.gov/artificial- intelligence (*website landing page*)	
User Control and Autonomy Framing	NA	"...We learned to prepare for new types of attacks that influence learning datasets, especially for AI systems that have automatic learning capabilities. To help ensure a similar experience [Tay episode] does not happen again, we developed technology such as advanced content filters and introduced supervisors for AI systems with automatic learning capabilities." (Unit 3 of Identify guiding principles for responsible AI module, Section: Novel Attacks)
Autonomy	NA	NA

User Control	NA	NA
Augmentation	NA	NA
Human Understanding	"All of the factors that influence meaningfulness contribute to the difficulty in modeling the interface between AI and humans. Developing systems that produce meaningful explanations need to account for both computational and human factors." "The tailoring of an explanation to user groups and individuals may not be static over time. As people gain experience with a task, what they consider a meaningful explanation will likely change." p.3	NA
Secure and Privacy Framing	NA	"AI will have implications on

		decision-making across industries, data security and privacy, and the skills people need to succeed in the workplace. As we look to this future, we must ask ourselves: How can we attain the benefits of AI while respecting privacy?"(Unit 3 of Identify guiding principles for responsible AI module, Section: Societal implications of AI)
Security and Safety (w.r.t data collection, processing, access, share, consent, data subject to AI decision making)	"By identifying and declaring knowledge limits, this practice safeguards answers so that a judgment is not provided when it may be inappropriate to do so." p.4	
The Other Framings	NA	NA

Ethical Dilemma and Moral Framing	NA	NA
Human Resource, Employment, Rights and Accessibility Framing	NA	NA
Fairness, Non-discrimination, and Justice Framing	” The Knowledge Limits Principle can increase trust in a system by preventing misleading, dangerous, or unjust decisions or outputs.” p.4	“Microsoft partnered with a large financial lending institution to develop a risk scoring system for loan approvals. We trained an existing industry model using the customer’s data. When we conducted an audit of the system, we discovered that while it only approved low-risk loans, all approved loans were for male borrowers. The training data reflected the fact that loan officers historically favor male borrowers—and inspecting the system allowed us to identify and address that

		bias before the system was deployed." (Unit 3 of Identify guiding principles for responsible AI module, Section: Biased Outcomes)
Accountability and AI Audits Framing	"This type of explanation assists with audits for compliance with regulations, safety standards, etc. The audience of the explanation may include a user who requires significant detail (e.g., a safety regulator) and the user interacting with the system (e.g., a developer). Examples may include the developer or auditor 263 of a self-driving car." p.4	"We believe that mitigating bias starts with people understanding the implications and limitations of AI predictions and recommendations. Ultimately, people should supplement AI decisions with sound human judgment and be held accountable for consequential decisions that affect others." (Unit 4 of Identify guiding principles for responsible AI module, Section: Fairness, Microsoft's first AI guiding principle)

		"Microsoft partnered with a large financial lending institution to develop a risk scoring system for loan approvals. We trained an existing industry model using the customer's data. When we conducted an audit of the system, we discovered that while it only approved low-risk loans, all approved loans were for male borrowers. The training data reflected the fact that loan officers historically favor male borrowers—and inspecting the system allowed us to identify and address that bias before the system was deployed." (Unit 3 of Identify guiding principles for responsible AI module, Section: Biased Outcomes)

Inclusion, Diversity, Solidarity, Protection of Cultural Differences and Whistleblowers Framings	NA	NA
AI Education, Science policy, and Public Awareness Framing	NA	"We anticipate these principles will evolve over time as we continue to learn and partner with customers, other tech companies, academics, civil society, and others on this issue. Review them in the summary and resources unit of this module." (Unit 3 of Identify guiding principles for responsible AI module, Section: Sensitive use-cases)
Responsible Research funding, Hidden AI Costs, Field Specific Deliberations Framing	NA	NA

Table 4

Examples of Identified Framings in the Institutional AI Ethics Principles and Guidelines Text Data

Discussion

As observed through the lens of framing theory, individual words (explained in the various academic framings and AI principles texts in this study) and their explanations or supporting (rhetorical speeches, assessment list or workshops) function as 'signs of priorities' within the AI principles and guidelines texts. This sort of institutional approach towards AI ethics and guidance providing documented resources with selective inclusions and intertwined framings confirms Goffman's (1974) and Entman's (1993) frame claims. The textual analysis of AI principles and guidelines corroborates Goffman's argument that there can be frames within frames and those connect to the 'person-role' formula directly and Entman's (1993) frame definition of selecting some aspects of a perceived reality and making them more salient in a communicating text.

Claiming to have put its responsible AI principles into action, Microsoft has shared its AI approach with partner and customer stories. The AI principles in action related website texts assure development and deployment of trustworthy AI in diverse sectors like auto and home insurance, banking, and telecommunications. Here, it was noted that instead of including all the framings of its AI principles, Microsoft prioritizes and selects framings specific to a field partner requirement. This prioritization and selection do not cover all the academic framings for every customer or societal partner. Also, the weight given to selected framings for a specific customer/partner while developing and deploying AI systems is not known. Thus, the final

outcomes in a social setting (like a bank) cannot be clearly stated either for the user (banker) at the setting (bank) or customers of that bank. Unlike Microsoft, each framing including *The Other Framings* (Diversity, Non-discrimination and Fairness, Accountability) was given equal treatment depth-wise in the EU's AI ethics document. AI HLEG also focused on *Societal and Environmental Well-being frame* which was not found in either of the other two AI principles text.

The diverse framings that came out of the AI principles and guidelines from different institutions are clearly stated to be legally un-binding but meets the rhetoric persuasion criteria to act as soft law for developer teams and leadership, explicitly supporting what Jobin et al. (2019) claimed in their AI ethics scoping review research. These diverse soft law perspectives and approaches contribute to the continuum of the global AI ethics, governance, and regulation discourse making that quite disperse and multi-meaning. Discourse being the production of knowledge through 'resources of communication' (Simons et al., 1976) like language, visual, audio-visual, new media entails meanings. These communications, their intended meanings, and interpretations are integral to societal functioning and socio-technical practices especially for new technology adoption in any society. These meanings shape and influence societal conduct (Hall, 2013).

As a highly transformative and proliferating advance technology - artificial intelligence- requires sound, solid, actionable, practical AI ethics and guidelines for *common-good* of a collective digital society. First requirement to realize this goal is formation of unified database for AI-specific ethics principles and guidelines or realization of some level of convergence at the global pioneers' level. This unification is no way a simple or time-bound process and asks for convergence (in meaning) from diverse AI players situated in different geographies, industries,

and in different societal roles. AI developer corporations like Microsoft have their own AI committees and offices for their AI development planning, strategy, problem resolution, and operationalization on societal grounds. The European Commission appointed its own AI expert group to operationalize its Trustworthy AI ethics guidelines through seven requirements and assessment lists which in its view would protect European citizens as the European authorities chart their AI progress and develop AI strategy for the coming fourth industrial AI revolution in the global arena. *"*Building on its reputation for safe and high-quality products, Europe's ethical approach to AI strengthens citizens' trust in digital development and aims at building a competitive advantage for European AI companies.*"* (European Commission, 2019, p.1). Acknowledging the absence of any such unified database for AI-specific ethics guidelines, Jobin et al., (2019) developed a protocol for their scoping review of the global AI ethics landscape. This approach helps in conducting AI ethics research studies and in highlighting the probable roadblocks in putting AI ethics into practice through AI downstream social applications. Some level of convergence (in AI principles' framings, intended meanings, social practice) at all levels of AI proliferation should help in developing TRUST in the transformational power of AI technology.

Utilizing moral philosophies, self-regulatory frameworks just as a brand or institutional communication strategy to situate an institution in legal, political, economic context will not help in resolving conflicts and tensions that arise when automated AI systems are deployed in actual social fields. Schnack (2020) in their work explained how scholarly-institutional steps towards improving the performance and interpretability of advanced artificial intelligence models in medical social applications like a brain disorder study led to further complexity in decision making and interpretability issues in the system. Warner and Sloan (2021) argued in their work

that *Explainability* should not be equated with *Transparency*. To address transparency and characterize its relation to explainability, they defined transparency for a regulatory purpose calling it ‘r-transparent’ (p.23). According to the researchers, a system is transparent for a regulatory purpose when regulators have an explanation, adequate for that purpose, of why it yields the predictions it does while as understood by computer scientists, a system is explainable if one can provide a human-understandable explanation of why it makes any particular prediction. Here, explainability of AI system remains relevant to transparency but turns out to be neither necessary nor sufficient for it. Emphasizing the importance of *transparency* and *interpretability* in social applications like healthcare and finance, where users’ *trust in AI systems* require rationale for the AI model's decision, Došilović et al. (2018) suggested developing criteria for AI systems’ interpretability, explainability, and trust while defining what trust, interpretability, comprehensibility, and explainability means with respect to advanced AI systems. In light of these points and scholarly understandings, agreement on principles and convergence on *AI for collective good* is valuable as that would serve a great deal for development of formal AI standards to adhere to and AI regulations to follow while putting AI ethics in practice from low risk to high-risk social scenarios.

Protecting few social values, attending to self-proclaimed priorities, or giving elite approved weight to a philosophy automatically compromises other values and goals that might be more important from a different or less-known perspective. For example, ‘fairness’ as an AI principle is important. From a political perspective there are spirited disagreements about what exactly constitutes fairness (Binns, 2017). From AI system designers’ perspective to design fairness-aware machine learning (a field that aims to enable algorithmic systems that are fair by design) a precise agreement on what it means to be fair is a must (Friedler et al., 2021). We need

to understand that the real world is structurally biased and thus produces structurally biased data that work as training data for AI systems. Working with AI training data, requires working with worldviews of AI stakeholders. These worldviews may be compatible or contradictory.

In the attempt to achieve convergence in AI ethics approaches, focusing on the 'framing' of ethical principles and guidelines will bring out insights for governing the scale and speed of AI's socio-technical progress in society. Communication theories and scholars' perspectives may help in demystifying the 'ivory tower' portrayal of black-box AI problems or tech ethics intellectualization that needs to be dealt with in practice.

In conclusion, responding to AI black box problems, risks, and concerns associated with advance AI approaches/ techniques through appointment of committees, AI expert groups, advisory councils and offices that are mandated to produce reports, recommendations, and guidelines for Ethical AI, is indicative of the intense and noteworthy efforts by diverse institutions. These efforts by AI developers, scholars, governments, and global authorities together with open stakeholders' participation are good starting points towards development and deployment of AI social applications for 'common-good'. The content of these myriad 'resources of communication' (Simons et al., 1976) requires attention and thorough analysis for attempting *convergence* as they shape our future global digital society and the field of AI ethics in a *soft* but *strong* way.

Study Limitation and Future Research

The proliferating AI technology is emerging, so is its connection to other technologies and societies it is operating in or will operate in future. As this process continues, the pioneer institutions like the European Commission, and NIST publish drafts of AI ethics principles and guidelines that get revised over time. This study is limited to the draft text accessed in

December 2021. Changes in the AI principles and guidelines post December 2021 are not accounted for in the textual analysis. This study selected three pioneers from different sectors, but they are just a few of many AI actors in their respective fields. Several national and international organizations, private sector corporations, professional associations, academic institutions, and non-profit organizations have been making efforts to address the societal concerns relating to AI technology by developing AI ethics guidelines, principles, frameworks, and reports with perspectives, suggestions, and action points that can be further analyzed for actual societal scenario or use-case specific implementation. There is a huge scope of research in the individual-level effect of frames as mentioned in the third area of framing research by Scheufele (1999). In case the AI principles are agreed upon broadly, they would miss out on recognizing the important and legitimate differences in values that exist across different social groups, diverse geographical populations and at individual levels. Holton & Boyd (2021) concluded in their research that the outcomes of the current AI systems are sometimes not human actors situated in socio-technical systems as AI developers have chosen. Institutional self-regulation with AI ethics principles and guidelines is just a starting point in this direction. Much needs to be done to explore the tensions that arise inevitably when these principles are put to practice (Whittlestone et al., 2019) in any social field. Thus, taking the formulation of AI ethics principles and guidelines as the first research area; the on-ground implementation and sector-wise or case-wise (also known as use cases or downstream application scenarios) conflict (or tension) resolution as the second research area; and finally, their impact study or effect at individual/team/group/society level as the third area of research would render *AI ethics approach for collective good* as potent ground for ethical action and fair implementation.

References

(2021, April 23). National Artificial Intelligence Initiative. https://www.ai.gov/

Akimoto, D. (2019). International regulation of "Lethal autonomous weapons systems" (LAWS): Paradigms of policy debate in Japan. *Asian Journal of Peacebuilding*, *7*(2), 311-332. https://doi.org/10.18588/201911.00a079

A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.),. (2021, March). *The FATE System: FAir, Transparent and Explainable Decision Making* [Paper presentation].

An external advisory council to help advance the responsible development of AI. (2019, March 26). Google. https://www.blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/

Arowolo, S. O. (2017). Understanding framing theory. *Mass Communication Theory*, *3*(6), 4.

ASME. (2021, January 11). National artificial intelligence initiative becomes law in FY2021 NDAA. The American Society of Mechanical Engineers - ASME. https://www.asme.org/government-relations/capitol-update/national-artificial-intelligence-initiative-becomes-law-in-fy2021-ndaa

Asplund, T. (2014). Climate change frames and frame formation: An analysis of climate change communication in the Swedish agricultural sector. https://doi.org/10.3384/diss.diva-105997

Bélisle-Pipon, J., Monteferrante, E., Roy, M., & Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01380-0

Benefo, E. O., Tingler, A., White, M., Cover, J., Torres, L., Broussard, C., Shirmohammadi, A., Pradhan, A. K., & Patra, D. (2022). Ethical, legal, social, and economic (ELSE) implications of artificial intelligence at a global level: A scientometrics approach. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00124-6

Binns, R. (2018). *Fairness in Machine Learning: Lessons from Political Philosophy* [Paper presentation]. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR 81:149-159, 2018.

Bradley, T. 2017, July 31. Facebook AI creates its own language in creepy preview of our potential future. Retrieved from Forbes: https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/#23250c7b292c

Bradley, T. (2021, December 10). *Facebook AI creates its own language in creepy preview of our potential future*. Forbes. https://www.forbes.com/sites/tonybradley/2017/07/31/facebook-ai-creates-its-own-language-in-creepy-preview-of-our-potential-future/

Caliskan, A. (2017). Beyond big data: What can we learn from AI models? *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. https://doi.org/10.1145/3128572.3140452

Caplar, N., Tacchella, S., & Birrer, S. (2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy*, *1*(6). https://doi.org/10.1038/s41550-017-0141

Carabantes, M. (2019). Black-box artificial intelligence: An epistemological and critical analysis. *AI & SOCIETY*, *35*(2), 309-317. https://doi.org/10.1007/s00146-019-00888-w

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20-23. https://doi.org/10.1038/538020a

Chien, S., Doyle, R., Davies, A., Jonsson, A., & Lorenz, R. (2006). The future of AI in space. *IEEE Intelligent Systems*, *21*(4), 64-69. https://doi.org/10.1109/mis.2006.79

Chong, D., & Druckman, J. N. (2007). A theory of framing and opinion formation in competitive elite environments. *Journal of Communication*, *57*(1), 99-118. https://doi.org/10.1111/j.1460-2466.2006.00331.x

Chong, D., & Druckman, J. N. (2007). Framing theory. Annual Review of Political Science, 10(1), 103-126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

Chuan, C., Tsai, W. S., & Cho, S. Y. (2019). Framing artificial intelligence in American newspapers. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3306618.3314285

Chui M., Manyika, J., Miremadi,M., Henke, N., Chung, R., Nel, P., Sankalp Malhotra, S.,. (2018). *Notes from the AI frontier: Insights from hundreds of use cases*. McKinsey & Company | Global management consulting. https://www.mckinsey.com/west-coast/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20AI%20frontier%20Applications%20and%20value%20of%20deep%20learning/Notes-from-the-AI-frontier-Insights-from-hundreds-of-use-cases-Discussion-paper.pdf

Coeckelbergh, M. (2020). AI ethics. https://doi.org/10.7551/mitpress/12549.001.0001

Crofts, P., & Van Rijswijk, H. (2020). Negotiating 'Evil': Google, project MAVEN and the corporate form. Law, Technology and Humans, 2(1), 75-90. https://doi.org/10.5204/lthj.v2i1.1313

Danaher, J. (2018). Toward an ethics of AI assistants: An initial framework. Philosophy & Technology, 31(4), 629-653. https://doi.org/10.1007/s13347-018-0317-3

D'Angelo, P. (2002). News framing as a Multiparadigmatic research program: A response to Entman. *Journal of Communication*, *52*(4), 870-888. https://doi.org/10.1111/j.1460-2466.2002.tb02578.x

de Greeff, J., de Boer, M. H.T. Hillerström, F. H.J., Bomhof, F., Jorritsma, W., Neerincx, M. (2021, March). *The FATE System: FAir, Transparent and Explainable Decision Making* [Paper presentation]. Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, CEUR Workshop Proceedings (CEUR-WS.org), Palo Alto, California, USA.

Deniz Susar D., & Aquaro V. (2019, April 3). Artificial intelligence | Proceedings of the 12th International Conference on theory and practice of electronic governance. ACM Other conferences.https://dl.acm.org/doi/abs/10.1145/3326365.3326420?casa_token=HWSN5pONpE0AAAAA:dAyrfIgCaIOYgIuu5LeJZJsZH0ASYD9-C2z7nCzUAhQGRB1IetYk05xp95dx2xFJSAKvtRQRkkXr

Druckman, J. N. (2001). The Implications of Framing Effects For Citizen Competence. Political Behavior, 23(3), 225-256. https://doi.org/10.1023/a:1015006907312

Endsley, M. R. (2017). Level of automation forms a key aspect of autonomy design. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 29-34. https://doi.org/10.1177/1555343417723432

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51-58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

European Commission, Directorate-General for Communications Networks, Content and Technology, (2019). Ethics guidelines for trustworthy AI, Publications Office. https://data.europa.eu/doi/10.2759/346720

Explainable AI. (2019). The Royal Society. https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf

F. K. Došilović, M. Brčić and N. Hlupić, "Explainable artificial intelligence: A survey," 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 0210-0215, doi: 10.23919/MIPRO.2018.8400040.

Foley, K., Ward, P., & McNaughton, D. (2019). Innovating qualitative framing analysis for purposes of media analysis within public health inquiry. *Qualitative Health Research*, *29*(12), 1810-1822. https://doi.org/10.1177/1049732319826559

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (Im)possibility of fairness. *Communications of the ACM*, *64*(4), 136-143. https://doi.org/10.1145/3433949

Goffman, E. (1974). Frame analysis: An essay on the organization of experience. Cambridge, MA: Harvard University Press.

Goffman, E. (1986). *Frame analysis: An essay on the organization of experience*. Northeastern University Press.

Google. (2019, March 26). An external advisory council to help advance the responsible development of AI. https://blog.google/technology/ai/external-advisory-council-help-advance-responsible-development-ai/

Google details formal review process for enforcing AI principles, plans external advisory group. (2018, December 18). 9to5Google. https://9to5google.com/2018/12/18/google-ai-principles-enforcement/

Google. (2020). Google AI. https://ai.google/static/documents/ai-principles-2020-progress-update.pdf

Google names external advisory council to guide artificial intelligence usage. (2019, March 27). 9to5Google. https://9to5google.com/2019/03/26/google-external-ai-advisory-council/

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Helmer, M., Schottdorf, M., Neef, A., & Battaglia, D. (2017). Author response: Gender bias in scholarly peer review. https://doi.org/10.7554/elife.21718.012

Hernandez, D., Cano, J., Silla, F., Calafate, C. T., & Cecilia, J. M. (2021). AI-enabled autonomous drones for fast climate change crisis assessment. *IEEE Internet of Things Journal*, 1-1. https://doi.org/10.1109/jiot.2021.3098379

Holland Michel, A. (2020). The black box unlocked: Predictability and Understandability in military AI. https://doi.org/10.37559/sectec/20/ai1

Holton, R., & Boyd, R. (2019). 'Where are the people? What are they doing? Why are they doing it?'(Mindell) situating artificial intelligence within a socio-technical framework. *Journal of Sociology*, *57*(2), 179-195. https://doi.org/10.1177/1440783319873046

How AI is helping with COVID-19 vaccine Rollout and tracking. (2020, December 17). AI Trends. https://www.aitrends.com/healthcare/how-ai-is-helping-with-covid-19-vaccine-rollout-and-tracking/

Ingram, K. (2020). AI and ethics: Shedding light on the black box. *The International Review of Information Ethics*, *28*. https://doi.org/10.29173/irie380

Innerarity, D. (2021). Making the black box society transparent. *AI & SOCIETY*, *36*(3), 975-981. https://doi.org/10.1007/s00146-020-01130-8

IPlytics. (2019, April). *Who is patenting AI technology?* IPlytics - The IP Intelligence tool. https://www.iplytics.com/wp-content/uploads/2019/03/IPlytics-AI-report.pdf

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2

Magazine, I. V. (2016, November 7). *Machine-learning algorithm quantifies gender bias in astronomy*. Scientific American. https://www.scientificamerican.com/article/machine-learning-algorithm-quantifies-gender-bias-in-astronomy/

Matthes, J. (2009). What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005. *Journalism & Mass Communication Quarterly*, *86*(2), 349-367. https://doi.org/10.1177/107769900908600206

Mhlanga, D. (2020). Industry 4.0 in finance: The impact of artificial intelligence (AI) on digital financial inclusion. *International Journal of Financial Studies*, *8*(3), 45. https://doi.org/10.3390/ijfs8030045

Miller, M. M., Andsager, J. L., & Riechert, B. P. (1998). Framing the candidates in presidential primaries: Issues and images in press releases and news coverage. *Journalism & Mass Communication Quarterly*, *75*(2), 312-324. https://doi.org/10.1177/107769909807500207

National Artificial Intelligence Initiative. (2021, April 23). https://www.ai.gov/

(n.d.). National Institute of Standards and Technology | NIST. https://www.nist.gov/system/files/documents/2019/06/03/iv-1.parker.nist-vcat-ostp-ai-june-2019.pdf

Obermeyer, Z., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287593

Office of the Chief Economist at the United States Patent and Trademark Office (USPTO). (n.d.). Inventing AI Tracing the diffusion of artificial intelligence with U.S. patents. www.uspto.gov/sites/default/files/documents/OCE-DH-AI.pdf

O'Neill, S., Williams, H. T., Kurz, T., Wiersma, B., & Boykoff, M. (2015). Dominant frames in legacy and social media coverage of the IPCC fifth assessment report. *Nature Climate Change*, *5*(4), 380-385. https://doi.org/10.1038/nclimate2535

Open University. (2013). *Representation: Cultural representations and signifying practices*. SAGE.

Our approach to responsible AI at Microsoft. (n.d.). https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5

Park, S. (2017). The fourth Industrial Revolution and implications for innovative cluster policies. *AI & SOCIETY*, *33*(3), 433-445. https://doi.org/10.1007/s00146-017-0777-5

Parviala, T. (2018). EU Entering the Era of AI: A qualitative Text analysis on the European Union's Policy on Artificial intelligence.http://www.diva-portal.org/smash/get/diva2:1280201/FULLTEXT01.pdf

Perez, C. C. (2019). *Invisible women: Data bias in a world designed for men*. Abrams.

Presidential memorandum for the Secretary of transportation. (2017). The White House – The White House. https://trumpwhitehouse.archives.gov/presidential-actions/presidential-memorandum-secretary-transportation/

Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, , Palo Alto, California, USA, http://ceur-ws.org/Vol-2846/paper35.pdf

Ratti, E., & Graves, M. (2022). Explainable machine learning practices: Opening another black box for reliable medical AI. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00141-z

Rich, E., Knight, K., & Kevin, K. (1991). *Artificial intelligence*. McGraw-Hill Science, Engineering & Mathematics.

Sætra, H. S., Coeckelbergh, M., & Danaher, J. (2021). The AI ethicist's dilemma: Fighting big tech by supporting big tech. AI and Ethics, 2(1), 15-27. https://doi.org/10.1007/s43681-021-00123-7

Schäfer, M. S., & O'Neill, S. (2017). Frame analysis in climate change communication. Oxford Research Encyclopedia of Climate Science. https://doi.org/10.1093/acrefore/9780190228620.013.487

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, *49*(1), 103-122. https://doi.org/10.1111/j.1460-2466.1999.tb02784.x

Schnack, H. (2020). Bias, noise, and interpretability in machine learning. *Machine Learning*, 307-328. https://doi.org/10.1016/b978-0-12-815739-8.00017-1

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea Teleoperators. https://doi.org/10.21236/ada057655

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B. (2022). Introduction: How to bridge the gap from ethics to practice. *Human-Centered AI*, 145-150. https://doi.org/10.1093/oso/9780192845290.003.0018

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics. *Journal of Database Management*, *31*(2), 74-87. https://doi.org/10.4018/jdm.2020040105

Simons, H. W. (1976). *Persuasion: Understanding, practice, and analysis*. Addison Wesley Publishing Company.

Simons H.W., MORREALE, J., and Gronbeck, B. (1976). Persuasion in Society. https://doi.org/10.4324/9780203933039

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607-1622. https://doi.org/10.1007/s13347-021-00477-0

Warner, R., & Sloan, R. H. (2021). Making artificial intelligence transparent: Fairness and the problem of proxy variables. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3764131

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3306618.3314289

Wilson, C. (2020). Artificial Intelligence and Warfare. In: Martellini, M., Trapp, R. (eds) 21st Century Prometheus. Springer, Cham. https://doi.org/10.1007/978-3-030-28285-1_7

Wilson, N.A. (2020). *Understanding the Battle for AI in Warfare through the Practices of Assemblage: A Case Study of Project Maven* [Doctoral dissertation].

Xu, Wei. (2019). Toward human-centered AI: A perspective from human-computer interaction. Interactions. 26. 42-46. 10.1145/3328485.

Appendix A

*Examples of Identified Framings in the Institutional AI Ethics Principles and Guidelines Text data (*High Level Expert Group on Artificial Intelligence set up by the European Commission*)*

TRUST Framings	Examples
Transparent and Comprehensible AI Framing	"An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention." p.13
Transparency	"This *(Transparency)* requirement is closely linked with the principle of explicability and encompasses transparency of elements relevant to an AI system: the data, the system and the business models." p.18
Explainability	"Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g., application areas of a system). "p.18
Interpretability	**NA**
Comprehensibility	**NA**
Reliable and Safe AI Framing	"...results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations. "p.17
Reliability	**NA**
Management practices directed towards Safety	**NA**
Public Reports of Problems/Failure/Misses/Future Plans	**NA**
Oversight Boards	**NA**
User Control and Autonomy Framing	"AI systems should support human autonomy and decision-making, as prescribed by the principle of respect for human

	autonomy. This requires that AI systems should both act as enablers to a democratic, flourishing and equitable society by supporting the user's agency and foster fundamental rights and allow for human oversight. "p.15
Autonomy	**NA**
User Control	**NA**
Augmentation	**NA**
Human understanding	"Whenever an AI system has a significant impact on people's lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (e.g., layperson, regulator or researcher)" p.18
Secure and Privacy Framing	"AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings…AI systems and the environments in which they operate must be safe and secure." p.12
Security and Safety (w.r.t data collection, processing, access, share, consent, data subject to AI decision making)	"In any given organization that handles individuals' data (whether someone is a user of the system or not), data protocols governing data access should be put in place. These protocols should outline who can access data and under which circumstances." p.17
The Other Framings	**NA**
Ethical Dilemma and Moral Framing	**NA**
Human Resource, Employment, Rights and Accessibility Framing	"In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited." p.20
Fairness, Non-discrimination, and Justice Framing	"The substantive dimension *(of fairness)* implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatization." p.12
Accountability and AI Audits Framing	"It necessitates that mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use…Auditability entails the enablement of

	the assessment of algorithms, data and design processes." p.19
Inclusion, Diversity, Solidarity, Protection of Cultural Differences and Whistleblowers Framings	"Equal respect for the moral worth and dignity of all human beings must be ensured… This also requires adequate respect for potentially vulnerable persons and groups, 21 such as workers, women, persons with disabilities, ethnic minorities, children, consumers or others at risk of exclusion." p.11
AI Education, Science policy, and Public Awareness Framing	"Beyond developing a set of rules, ensuring Trustworthy AI requires us to build and maintain an ethical culture and mind-set through public debate, education and practical learning." p.9
Responsible Research funding, Hidden AI Costs, Field Specific Deliberations Framing	**NA**