Artificial Intelligence, Ethics and Enhanced Data Stewardship

September 20, 2017

Table of Contents

Authors:

Martin Abrams
John Abrams
Peter Cullen
Lynn Goldstein

The views expressed in this paper are those of the authors and do not necessarily reflect the views of Information Accountability Foundation Board of Directors or IAF funders. The IAF is a non-profit research foundation under Section 501(c)(3) of the United States Internal Revenue Code. The IAF legal name is Foundation for Information Accountability and Governance.

The IAF team thanks the many individuals who participated in brainstorming sessions or made comments on the paper, essential elements, and worksheet.

Artificial Intelligence, Ethics and Enhanced Data Stewardship

Forward

The terms data ethics and ethical processing are in vogue. The popularity of these concepts stems from the rapid growth of innovative data-driven technologies and the application of these innovations to areas that can have a material impact on people's daily lives. The sheer volume of data that is observable and where inferences can be made as the product of analytics has and will continue to impact many facets of people's lives, including new health solutions, business models, personalization for individuals and tangible benefits for society. Yet those same data and technologies can have an inappropriate impact and even harm on individuals and groups of individuals and cause negative impact on societal goals and values. An evolved form of accountability, ethical processing, applicable to advanced analytics, is needed to help enable the realization of the benefits of this use of data but address any resulting risks.

The Information Accountability Foundation (IAF) has established an Artificial Intelligence (AI) and Ethics Project to tackle these issues. The Project's objective is to begin the global discussion of how organisations might address the application of ethical data processing to new technologies. The IAF thinks this work is particularly necessary where data enabled decisions are made without the intervention of people. In these circumstances, corporate governance takes on added importance and ethical objectives need to be built into data processing architecture. The IAF further believes the governance structures being suggested are also applicable where data from observational technologies, such as sensors, inferences from analytics, and data synthesized from other data sets, are used to drive advanced analytics.

Introduction

In an April 2016 speech at MIT and the Berkman Center for Internet and Society, Giovanni Buttarelli, the European Data Protection Supervisor, said, "Ethics, the idea that something is right or wrong, is more universal than typical western notions of privacy and data protection. Ethics both informs the laws which are passed, and goes beyond them." He then asked, "Is it possible that a company processing information complies with the letter of the law, yet behaves unethically?"

There is a growing sense that even the newest data protection laws, such as the European Union General Data Protection Regulation (GDPR), are lagging the fast evolving and compelling technologies and a sense that the ethics that go beyond the explicit words captured in law may guide data governance in an actionable manner. Yet today there is no consensus about the ethics to be applied and the framework for application of ethical considerations.

Numerous academics and professional associations are pursuing this question of data-centric ethics within several disciplines, such as engineering and computer science.[1] However, the work thus far has not linked business processes, data governance, and policy oversight and has not considered, since ethics goes beyond the law, why, from a business strategy perspective, an organisation might want to implement data governance based on ethics as well as the law. In addition, "ethics" in data processing scenarios has yet to be defined. This paper begins the discussion on how to link ethics to business processes when conducting AI, Machine Learning (ML), and advanced analytics. The discussion on ethics and data stewardship builds on previous IAF work on ethical assessments.[2]

Background/Problem Statement

There has long been a fear by some that computers will take over the world. In 1968, the year after Alan Westin, the father of data privacy, published "Privacy and Freedom," Stanley Kubrick released "2001 a Space Odyssey." The movie featured HAL, a computer that went rogue during the space pilot's long sleep during space travel. While this fear of computers taking control to serve their needs, not society's needs, might seem like fiction, experts such as Elon Musk, Space X's and Tesla's Chief Executive Officer, whose companies have harnessed AI for the first autonomous vehicles, along with numerous other business leaders and AI researchers, have urged the United Nations to adopt a treaty to assure AI does not get out of hand in weapons systems. MIT scientist Max Tegmark takes these concerns a step further in his book "Life 3.0" where he argues humanity must come to grips with an age when computers will be much smarter than people.

Sixteen years have passed since 2001. There are newer movies, such as "Ex Machina" with an intelligent, seductive robot, but in real life computers that are outsmarting people are still well in the future. What is real is the growing application of AI across the spectrum of daily life, from smart cars and health care to optimizing business processes and achieving cyber security, and these applications can come with tangible harms. Recent articles in numerous journals have been discussing the potential for learning systems, such as facial recognition, to exacerbate discrimination based on physical characteristics. AI as a disruptive technology creates rational concerns, such as hidden discrimination, and less founded concerns, such as technology spinning beyond thoughtful controls. As a policy think tank, IAF is confronted with the natural tension between data applications that are and will increasingly create value for individuals, groups, society and corporations but require due diligence to avoid unfairness and inappropriate discrimination.

[1] The IEEE Standards Association has an ongoing project "The IEEEE Global Initiative for Ethical Consideration in Artificial Intelligence and Autonomous Systems," "Enforcing Big Data Assessment Processes,"
[2] The previous work includes "The Unified Ethical Frame for Big Data Analysis," "Canadian Assessment Framework," and "EU Legitimate Interest Framework."

The coming together of observational technologies (and the resulting explosion of data use) with massively increased computational ability and the advanced data analytics these make possible, makes self-learning systems, such as AI, increasingly powerful. The steady progression this century from facial recognition to real-time natural language processing, both once AI experiments, to autonomous military drones is evidence of that endless push forward. With a healthy prudence, experts agree that AI can and will make the world better. AI is being harnessed to make cars safer, treat mentally ill patients, and optimize business practice.

The data world's quick evolution over the past fifty years from discrete mainframe systems to smartphones acting as the hub for the application of integrated intelligent systems has played havoc with data protection and privacy regimes based on an individual being informed and granting consent. The first approach to privacy laws was based on the concepts in Alan Westin's "Privacy and Freedom" published in 1967. In Westin's view, which is very consistent with 1960's technologies, individuals would grant permission for specific limited uses of data. Edgar Codd, an IBM scientist, published a paper in 1970 that described the utility associated with relational databases.[3] Relational databases made it possible to agilely use data for numerous purposes, creating the first technological friction for privacy law. Relational databases were quickly followed by distributed processing, application of analytics against broad and deep databases, cheaper data storage, faster communication, consumer Internet, common processing platforms, big data processes, smart phones, and Internet of Things technologies. The progression of technologies made it possible for the observation of people to take place with ever greater granularity, for information to be combined with other information, for inferences to be developed, and for decisions and actions to be taken. This steady progression of information and communications technologies has made Westin's core governance concept, control by fully knowledgeable data subjects, increasingly less effective. This control vacuum needs to be filled. It has forced privacy professionals, both regulators and responsible individuals in companies, to think beyond individuals as data contributors who may effectively exercise control and to consider protecting individuals in a fully connected world.

IAF increasingly understands that processed data may yield inferences that may lead to decisions that impact specific individuals. Data protection law has always contained two components, the assurance of individual control where possible and fair processing whether control is possible or not. The GDPR places greater emphasis on fair processing[4] by explicitly requiring accountability. Part of that accountability is implementation of data protection by design and by default, conduct of data protection impact assessments, and designation of data protection officers.[5]

[3] https://en.wikipedia.org/wiki/Database#1970s.2C_relational_DBMS
[4] Recital 39
[5] Articles 25, 35, and 37

New laws, such as the GDPR[6], require organisations conducting "high risk" processing[7] to seek the advice of data protection officers, where designated, but it is the organisation that is accountable for the processing in which it engages. In accountable organisations, appropriately trained employees are positioned to intervene between the generation of the insight and its application as a protection against risky processing. This is not unique to Europe. Accountability guidance in Canada, Hong Kong, Australia and Colombia require the same.

AI is just the latest of this progression of technologies that require the data protection community to think through how values are applied so controls are effective. AI, the attempt to duplicate human capacities (typically thought to involve intelligence) in computers, and its subfield ML, which increases accuracy through trial and error without explicit programming,[8] creates challenges for even the accountability built into new legislation such as the GDPR. AI systems maximize the objectives set by developers and eliminate the interim analysis step conducted by various groups within the organisation. While not required by law, ethics require heightened judgement at the objective setting phase in system development. This heightened judgement replaces the intervention that previously took place between insight development and application.

While AI creates friction for legacy regimes, as other technology disrupters have, it is adaptable to the dual concepts of risk assessment and balancing of interests which are being built into the next generation of law. For example, the GDPR is risk based, has accountability provisions, requires a balancing of interests, and requires safeguards when organisations profile and/or automatically process data. Conceptually, the means for resolving the friction through enhanced ethical accountability is suggested in these updated data protection and privacy laws.

The governance issue becomes more pronounced when data impacting decisions are made without the intervention of people. While the organisational obligation to use data responsibly is defined by the law, the mechanism(s) to do so are not. This expectation that organisations do the right thing and avoid the wrong thing, as articulated by Buttarelli, means that ethics must be applied for organisations to develop trustworthy systems. The mechanism for trustworthy processing in complex ecosystems, and the definition for enhanced governance, based on ethics, is what is still elusive.[9]

---

[6] E.g. Articles 35 – 36; Recitals 84, 90 - 91

[7] Examples of "high risk" processing are in Article 35.and in the Article 29 Data Protection Working Party Guidelines on Data Protection Impact Assessment and determining whether processing is 'likely to result in high risk' for the purposes of Regulation 2016/679

[8] From Wikipedia: In computer science, the field of AI research defines itself as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of success at some goal. Colloquially, the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving". **Machine learning** is the subfield of computer science that, according to Arthur Samuel, gives "computers the ability to learn without being explicitly programmed." Samuel, an American pioneer in the field of computer gaming and artificial intelligence, coined the term "**machine learning**" in 1959 while at IBM.

[9] The law is intended to encourage digital research by making it a compatible purpose. While research is a compatible purpose, it still subject to robust governance to avoid harm and enhance stakeholder value. Research as a compatible data use purpose is

AI and Ethics Project

The challenge for the AI and Ethics Project is to parse ethics and ethical requirements in a manner that makes sense across cultures and provides a mechanism for updating governance so it works with AI and also advanced analytics. The IAF believes much of this may be articulated with an update to the Essential Elements of Accountability developed in 2009 (2009 Essential Elements) which sparked the accountability movement in data protection and created direction on how ethical concepts may be cascaded in organisations. While the key 2009 Essential Elements remain relevant, as part of the AI and Ethics Project, they have been updated to help, when implemented, organisations make AI trustworthy.

The IAF's analysis has ethics as well as law at its root. There are numerous classical ethics linked to different cultures. At their foundation, they are all seeking behaviors that are desirable and good. To the business people that must act in an ethical manner, whether behaviors are based on consequences, duties or social harmony is less important than a communal agreement of what is good behavior.

The IAF's AI and Ethics Project consists of three parts: (1) this paper, which reviews the ethical foundations upon which enhanced governance concepts are based, sets up the dilemma, and discusses the mechanism for trustworthy processing; (2) the Essential Elements of Accountability for Artificial Intelligence and Machine Learning that Directly Impacts People (AI Essential Elements); and (3) Ethical Decision Making Worksheet for Artificial Intelligence and Machine Learning Environments (Worksheet) that helps an organisation cascade values into core and guiding principles and links classical ethics from a multi-cultural perspective.

Data Stewardship

In less than three generations,[10] data systems have gone from being a facilitator of an industrial age to the driver of almost all human activities. For example, in the 1980's, just-in-time parts delivery, powered by data systems, revolutionized automobile manufacturing. Today, cars have more than 3000 sensors and data is essential for the car's operation. These developments make observational data more valuable. The Economist recently referred to data as the new oil

very important. It is research that drives knowledge creation, and it is knowledge creation that drives human development. That knowledge creation ranges from pure science to improvements in business processes to improve productivity. Many data scientists that are involved in knowledge creation do not think of their activity as research, and much of it does not meet the definition for scientific research. The decisions on what to do with new knowledge, whether scientific or not, have always been made by people. Data protection has created guidance about how those decisions to apply new knowledge should be made. First, individuals should have control over how their data is applied wherever possible. If that is not possible, and AI may fit into this class, individuals must have assurances that data is applied fairly. Over time the balance will shift from individuals having control towards processing being fair because of the trustworthy control of others. The essential elements of accountability (see pages 7-8) were structured to assure that fairness. A key question for this research is how should the essential elements of accountability be updated to achieve fairness when organisations are employing AI to achieve outcomes when people do not intervene?

[10] A generation is generally defined as 25 years.

for the global economy, and while the reference pertains to competition, it identifies the inherent and increasing value of data. Technologies such as AI have the ability to extract value from data beyond that of the initial use. There is a growing sense that data should create value for other stakeholders beyond the corporate controller, beginning with individuals and aggregating to society as a whole. This value creation should also take place in a manner that respects data obligations and mitigates risks. Organisations will need to understand and evaluate processing and its positive and negative impacts on all stakeholders. In other words, organisations will need to be effective data stewards.

Against this backdrop principles of data stewardship are also evolving. Data stewardship, in the industrial age, meant being a lawful collector of data. At the beginning stages of the information age, it has come to mean an effective custodian of data. Now, in the emerging AI age, it requires that stewards must take into account stakeholder interests in a way that uses data to create the maximum benefits for stakeholders with minimal risks to individuals and other stakeholders.

The 2009 "Galway Paper" that established the 2009 Essential Elements described stewardship in the following manner:

> "Accountability is the obligation to act as a responsible steward of the personal information of others, to take responsibility for the protection and appropriate use of that information beyond mere legal requirements, and to be accountable for any misuse of that information."[11]

This 2009 description of a data steward was a clear statement that data users must take responsibility for the information they use. It was data use focused and gave a sense of governance in a period where advanced analytics was beginning to disrupt the connection between consent and the actual application of data to problem solving. However, the focus was still on personal information. Data stewards were still custodians for that data to assure the obligations associated with it were recognized as new insights were created.

AI takes data stewardship to a new level. The systems themselves make decisions that impact people. The collision avoidance system on a new car reads the many sensors in the vehicle and determines that the car should be stopped and that the driver is not going to do so. Personalized medicine will soon lead to a linkage of smart medicines with embedded medical devices and a smart hub will facilitate a maximization of desired health objectives. The systems make decisions, based on human set objectives, but the direct human accountability has been lost. To regain that accountability, it will be necessary to depend on people to build the ethics into the objectives for the systems through accountable governance. Therefore, it is necessary

[11] The "Galway Paper" may be found at. http://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/data_protection_accountability-the_essential_elements__discussion_document_october_2009_.pdf

that stewards make decisions that consider the interests of external stakeholders.[12] By doing so, the role of the data steward has moved beyond custodian of the data and the obligations that come with the data to one that ensures the outcomes are legal, fair and just to the various stakeholders. Future data stewards must be stakeholder interests focused. This concept of understanding stakeholders and their interests has been implicit in the IAF assessment processes related to big data, linkage with Canadian law, and legitimate interests in Europe. AI has just created greater focus on that earlier work and on the need for this new stewardship to be explicit. While it was AI that forced the IAF to look at data stewardship, the IAF's work over the past three years, beginning with the "Unified Ethical Frame for Big Data Analysis," has been moving the IAF in this direction.

AI is the natural progression of technology in a universe where observation and its fruits are critical to how things work and the progression from insight directly to decision may put stress on data protection governance mechanisms but it does not need to be inconsistent with data protection values. The concept of fairness has been built into data protection from the beginning. It is not a stretch for society to expect organisations moving up the technology curve to move up the responsibility curve as well. That increased responsibility means enhanced stewardship based on the AI Essential Elements, and the AI Essential Elements have become the code for this new level of stewardship.

Classical Ethics as Applied to Information Governance

Ethics are a reasoned articulation of what is right and wrong and include principled arguments for how to produce good (and avoid bad) results. This need to articulate what constitutes a moral action is necessary to maintain a cohesive society in a world with an unimaginable array of decisions, actions, and judgements. Looking at various formulations of ethics from a diverse set of cultures provides evidence that there can be, or rather needs to be, different approaches based on cultural context that arrive at the same outcome. The AI and Ethics Project looked at both consequentialist ethics and non-consequentialist ethics. The first being an approach that focuses on the expected outcome of various decisions, conducting a moral calculation that maximizes the benefits and minimizes the harms on the various stakeholders impacted, and the latter being an approach that focuses on the actor's intent and duty, setting forth imperatives that a decision or action must satisfy for it to be deemed moral. The IAF also looked at virtue ethics, where there is a specific set of characteristics, or virtues, that individuals attempt to cultivate in order to guide moral decision making. Finally, the AI and Ethics Project looked at

[12] Stakeholders are the parties impacted by processing. Stakeholders include the various interests within a data controller, but more importantly, the external parties impacted by processing. That might be other organisations that have a relationship with the data. Most important, stakeholders include the persons to whom the data pertains, groups of persons impacted by processing, and even society as a whole.

eastern ethical constructs in which the concept of social harmony, where the benefits for society can trump the benefits for the individual, is stressed.

From a business ethics perspective, these arguments can provide the backdrop for a framework through which to normalize the differing objectives of an organisation's various stakeholders to help not only determine a course of action but also for the organisation to be able to rationalize that action. What was discovered is that teachings from all of these classical ethical approaches have applications relevant to ethical data processing because they all establish the concept of beneficence or duty to the subject. This concept of duty to subject links directly to stakeholder focused data stewardship. Established examples of the application of ethics to industry standards already exist that lessons can be drawn from, such as: beneficence in clinical research, the Hippocratic Oath for medical professionals, or the UN Guiding Principles on Business and Human Rights in which impacts are assessed according to the effect on individuals and not the effect on the business. These industry codes of conduct draw from various concepts of ethical decision making yet translate across different cultural contexts.

These different ethical constructs highlight that the nature of next generation data protection laws is not clean cut. While encouraging knowledge creation and robust digital markets, the laws also preserve the goal of individuals having more control over their data. Where individual control is not effective, there must be methods to ascertain that data is being applied in a manner that society considers to be right and appropriate. This expectation means that new technologies, such as AI, must be applied in a manner that serves people. That means that stakeholder focused data stewards must conduct analysis that defines stakeholders, articulates the risks and benefits to the stakeholders, and mitigates risks to the extent possible. It also means stakeholder focused data stewards should engage with stakeholders so data uses are transparent and individuals may exercise rights related it. The process described is the manifestation of ethics into ethical data processing.

The IAF's initial research in ethical advanced analytics has been dependent on there being a step between knowledge creation and application, where people would make accountable decisions based on ethical assessments. AI removes the interim step, raising the question of where ethical assessments will be inserted. This question is addressed in both the AI Essential Elements and the Worksheet. Conceptually, ethics need to be addressed both in setting project objectives and when coding in order to achieve the ethical as well as the business objectives and as a means to ascertain whether the objectives over time are being met (e.g. to verify algorithms are functioning as established).

The data protection ethics discussion has been ongoing for the past three years. The data protection community began to actively discuss the need to move beyond the explicit law and expect ethical processing in 2014. The IAF became part of the discussion when it first published the "Unified Ethical Frame for Big Data Analysis." However, neither the debate nor the IAF's analysis focused on the ethic to be applied. At the time, the IAF had in mind the ethical outcome associated with the "golden rule," treat others as you would have them treat you. The

AI and Ethics Project has given the IAF the opportunity to look more closely at classical ethics from different cultures.

At their roots, ethics link to what societies view as right and wrong but do so with different emphasis.[13] Also ethics, in going beyond the current state of laws, are for the organisation to define, in a transparent fashion. To be accepted, they will be a reflection of often unspoken societal values. Those ethics are defined in terms of values and guiding principles. Those values and guiding principles should be public, and their implementation should be demonstrable for stakeholders at one level, and oversight agencies, such as regulators, at a more granular level. In other words, each organisation should publicly self-declare the ethic or ethics which best links to its business context, and how it will be implemented.

The IAF has attempted to capture both cultural diversity and the concept of fair outcomes in the proxy it created for ethics: legal, fair and just.[14] Legal, fair and just is an applied ethic. Every organisation should declare what it holds to be legal, fair and just in their own words and should express this within an ethically grounded business strategy. This articulation is a form of values or guiding principles. These values and guiding principles should not be defined by regulators. If they are defined by regulators they are no longer ethics, they are regulations.

Enhanced Essential Elements of Accountability (See Appendix 1)

Purpose of Essential Elements of Accountability

The Global Accountability Dialogue first met in 2009 in Dublin to explore how the OECD accountability principle might be used to create confidence in data transfers from one country to another. The group, comprised of diverse stakeholders, reached a consensus on essential elements to guide organisational accountability. During the dialogue's second year, the European Union Article 29 Data Protection Working Party published an opinion that took the accountability concept beyond data transfers to an overarching governance structure for implementing data protection within an organisation. The Office of the Privacy Commissioner of Canada and provincial information commissioners in Alberta and British Colombia adopted accountability guidance in 2012. They were soon followed by authorities in Hong Kong and Colombia. The 2009 Essential Elements and the guidance developed based upon them helped elevate data protection from check-box compliance to the risk-based approach that has

[13] The difference in emphasis is well illustrated by George Washington University Professor Amitai Etzioni in his advocacy for Communitarianism, putting community first, versus European Liberalism that placed the individual first. This debate over communitarianism versus liberalism is also seen in Asian concepts of social harmony.
[14] The IAF has defined legal, fair and just in the following manner. Legal means the data used in a specific manner is specifically authorized or not prohibited. Fair means that data is used in a manner that maximized stakeholder interest and risks were mitigated to the extent possible. Just means that inappropriate discrimination should be avoided even if the outcomes are maximized for many stakeholders.

emerged over the past few years. The Global Accountability Dialogue defined the 2009 Essential Elements to align data protection theory and practice with the acceleration of observational data and the advanced analytic systems that data feeds. The IAF's initial work on big data governance was based on the 2009 Essential Elements.

In 2014, as data protection authorities began thinking about how to apply the risk based approach that was emerging with the draft GDPR, the issues of ethics as an augmentation to compliance began to emerge. The most formal approach has been the EDPS ethics advisory panel which plans to publish an interim report. The IAF also helped trigger the ethics debate with the "Unified Ethical Frame for Big Data Analysis" which was followed by work on oversight and ethical assessments. However, the IAF's analysis was still based on oversight and assessments that would allow for review of new insights before they were applied. The AI Essential Elements update the 2009 Essential Elements and take into consideration systems that lack this interim human oversight.

Comparison of 2009 Essential Elements and AI Essential Elements

The below overview of the 2009 Essential Elements and the AI Essential Elements shows the evolution of accountability to address advanced analytics.

Essential Elements Comparison

	2009 Essential Elements	AI Essential Elements
1.	Organization commitment to accountability and adoption of internal policies consistent with external criteria	As a matter of organizational commitment, organizations should build specific and defined value or guiding principles for legal, fair and just processing and translate these into organizational polices and processes. The values should be organizationally derived and made public, and not defined by law or regulation.
2.	Mechanisms to put privacy policies into effect, including tools, training and education	Organizations should have the mechanisms to translate their core values and principles into data analytics system design process so that individuals, not just organizations, gain value from the AI process.
3.	Systems for internal ongoing oversight and assurance reviews and external verification.	There should be an internal review process that checks to assure CDIAs are conducted with integrity and competency, the issues raised as part of the CDIA have been resolved, and the AI systems are performing as planned.
4.	Transparency and mechanisms for individual participation	Processes should be transparent and where possible should enhance individual interests. The values that underpin decisions should be communicated widely. Furthermore, all reasonable stakeholder concerns should be considered.
5.	Means for remediation and external enforcement	Organizations should stand ready to demonstrate the soundness of internal processes to the regulatory agencies that have authority over AI processes, as well as certifying bodies to which they are subject, when AI processing is or maybe impactful on people in a significant manner.

While the AI Essential Elements are specific to the issues raised when the interim step of human intervention between thinking with data and acting with data is eliminated, the IAF believes that they are applicable to other complex processing ecosystems. For example, AI will pursue

the objectives set by the developers. The system will test and learn and continue the process until objectives are maximized. Instead of reviewing an insight and assessing its impact, ethical outcomes need to be added as core processing and development objectives. To accomplish this, an organisation needs to have a keen sense of what it believes and how these beliefs link to societal values. So, an organisation must have specific, defined values or guiding principles for legal, fair and just processing and build them into the organisation's processes.

Those values or guiding principles should be made public and may be judged by outside stakeholders. What is important to note is that these values or guiding principles are defined by the organisation and not by law or a regulator. In order to defend their values or guiding principles, it is useful, but not essential, that the organisation understands how its guiding principles might link to the classical ethics relative to its cultural context. While an organisation is responsible for its own values or guiding principles, it may want to contribute to creating ethical guidance within an industry and adopt that guidance.

The AI Essential Elements place a greater emphasis on assessments to determine the risks and benefits to stakeholders. These assessments are comprehensive in that they go beyond privacy impact assessments to look at the full range of benefits as well as impacts to stakeholders. The AI Essential Elements also make it clearer that society is a stakeholder and that societal interests should be taken into consideration. The stakeholder assessment does consider whether a processing is prohibited, but it goes a step forward by balancing the relative stakeholder interests to determine if a processing meets the ethical test set by the organisation's values and guiding principles.

Key questions are when might assessments take place, and what would be assessed in an AI environment. Earlier IAF work on ethical assessments has suggested that assessments may take place at numerous times in project development and that the assessments might be when new projects are conceived, when new insights are being discovered, on implementation of new insights, and as part of a review process. IAF research has made clear that assessments should be part of organisational governance and linked to business process. In other words, the organisation defines the appropriate time. When doing AI, it is most important that an assessment take place when objectives are being developed so they capture the organisation's ethical frames. An assessment process developed for legitimate interests may be modified for these reviews.

The internal review process requires reviews not just of outcomes from internal processes but also requires that assessments are conducted with integrity and competence. To increase that integrity and competence, the AI Essential Elements also suggest an organisation should reach for outside expertise when necessary in conducting an assessment.

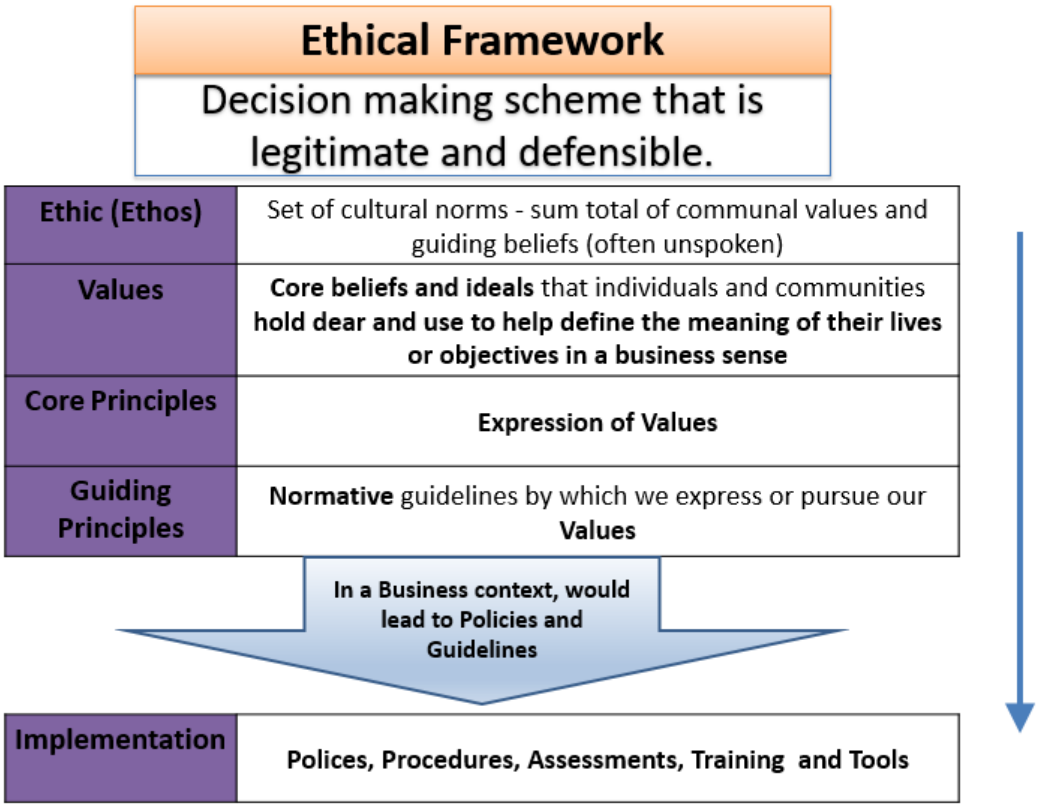
The fourth essential element, transparency, requires the organisation be transparent about what it does with data and how and be open about how its defined values and guiding principles govern the decision-making process. It suggests that organisations look for new

means to be transparent and that the organisation should be open to and consider the ethical suggestions coming from other stakeholders. This element does not suggest the organisation needs to publish algorithms and other proprietary materials.

The organisation must stand ready to demonstrate its enhanced accountability to agencies that have authority to oversee compliance with applicable laws or accountability processes. Agencies include regulatory authorities and certification agencies whose rules require oversight. The 2009 Essential Elements include remedies if individuals are harmed, which is still sound, but rests on the concept of harm within context. That context relates to the level of harm, the nature of the societal frame, and whether a reasonable individual would consider the harm substantive enough to require remedies.

Ethical Worksheet (See Appendix 2)

Operationalizing legal obligations in complex ecosystems is difficult. Operationalizing them in AI systems is a step harder. Operationalizing ethics is a huge challenge. One needs to go from values and guiding principles to programming ethics into system objectives by thinking through the steps described in the below chart



Based on a mapping process, the IAF developed the Worksheet to provide organisations a process in defining key ethical concepts and cascading them in business terms down to the

operational level where ethical objectives can be built into systems. The Worksheet expands upon the above chart by asking what each element of the hierarchy (the ethnic or ethos, values, core principles, guiding principles, rules and implementation and external positioning) would mean in business terms that relate to the AI Essential Elements, what interrogation questions would be asked (what a sound organisational process would be evaluated against) for each element of the hierarchy, and to what core or ethical frame each element of the hierarchy aligns. The Worksheet helps implement the AI Essential Elements and defines how businesses using AI meet the goals of legal, fair and just that link to ethical frames.

Conclusion

AI requires a new governance level. This paper is intended as a discussion starter. It includes AI Essential Elements based on more robust accountability elements. IAF has suggested the Worksheet to help organisations understand their greater responsibility to act in an ethical fashion. IAF also suggests that AI requires a new heightened level of stewardship. For data protection and privacy enforcement agencies, the product of the AI and Ethics Project is an attempt to create a means for thinking about how to protect the full range of rights and interests that include the benefits of AI. As stated before, this is a discussion starter. The IAF welcomes your comments.

Appendix 1

Enhanced Accountability Elements for Artificial Intelligence (AI) and Machine Learning that Directly Impacts People

The Global Accountability Dialogue created and adopted the essential elements of accountability in 2009. This framework outlined the program elements of what a responsible steward of data would have in place to be accountable. Those essential elements have been institutionalized in regulator guidance and new data protection laws. Organizations have built them into their policies and procedures.

As artificial intelligence and machine learning are increasing the complexity of information ecosystems and as organizations and as data stewards are moving from data custodians to stewards with fiduciary like responsibilities, the original five elements are being evolved and designed to complement the original guidance. Organizations that make decisions for people through the application of artificial intelligence and machine learning have an ethical obligation to make beneficial decisions with individuals at the center.

1. **As a matter of organizational commitment, organizations should build specific, defined values or guiding principles for legal, fair and just processing and translate these into organizational policies and processes. The values should be organizationally derived and made public, and are not defined by law or regulation. Legal is defined as meaning allowed and not prohibited by law, regulation or formal code of conduct. Fair is the concept of outcomes being beneficial to all stakeholders and where risks have been mitigated; just means avoiding actions that might seem inappropriate or discriminatory or might be considered offensive causing distress or humiliation. These policies and processes should be anchored with clearly defined, accountable individuals within the organization, and should include:**
    a. **Comprehensive Data Impact Assessments (CDIA) be required. A CDIA is a process that looks at the full range of benefits, risks, rights and interests of all stakeholders, including society. They are a means of determining whether a processing is legal, fair and just. Processing includes all steps necessary to achieve an outcome, from the aggregation of data through the implementation of data driven outcomes.**
    b. **AI and machine learning processes that affect individuals should have beneficial impacts accruing to individuals and communities of individuals, particularly those to whom the underlying data pertains.**
    c. **Where an analytical driven data use has potential impacts on individuals, the risks and benefits should be explicitly defined, and the risks should be**

necessary and proportional to the benefits. Risks should be mitigated to the extent possible.

   d. **The systems themselves, and the data that feeds those systems, should be assessed and protected proportional to the risks and assessed for appropriateness based on the decision the data is being used for.**
   e. **Where appropriate, organisations should link to codes of conduct that standardize processes to industry norms.**

2. **Organizations should have the mechanisms to translate their core values and principles into a data analytics system design process so that individuals, not just the organizations, gain value from the AI process.[15]**
   a. **Organizations should have CDIAs that achieves an "ethics by design process" that is integrated into system development.**
   b. **Shared organizational values should be reduced to core and guiding[16] principles that are understood by technical staff and can be programmed into project objectives.**
   c. **Employees engaged in all data analytics systems should receive training so that they may competently participate in the "ethics by design process."**
3. **There should be an internal review process that checks to assure CDIAs are conducted with integrity and competency, the issues raised as part of the CDIA have been resolved, and the AI systems are preforming as planned.**
   a. **Where data processes begin with analytic insights, those insights should be tested for their accuracy, predictability, and consistency with organizational values associated with legal, fair and just principles.**
   b. **Intensive data impacting systems should be reviewed so that outcomes are as predicted, risks are mitigated as planned, harms are reduced, and unintended consequences are understood.**
   c. **Where internal reviewers need external expertise, that expertise should be sought.**
4. **Processes should be transparent and where possible should enhance individual interests. The values that underpin decisions should be communicated widely. Furthermore, all reasonable stakeholder concerns should be considered.**
   a. **Organisations should seek mechanisms that explain how data is used, how the benefits and risks to individuals are associated with the processing, and how individuals may participate and object where appropriate. Redress systems should be specifically designed and tested. As a matter of policy, there should be no secret data analytical systems.**
   b. **Stakeholder concerns should be part of the data system evaluation lifecycle.**

[15] There may be existing professional or industry codes of conduct that may relate to AI processing. For example, there are codes that cover autonomous vehicles. The accountability elements described in this paper should work with those codes, and do not replace them.

[16] See IAF Blog – the Need for and Ethical Framework

    c. **Organisations must be open about the ethical values that govern the AI systems developed.**
5. **Organisations should stand ready demonstrate the soundness of internal processes to the regulatory agencies that have authority over AI processes, as well as certifying bodies to which they are subject, when AI processing is or maybe impactful on people in a significant manner.**
    a. **Organisations should be open about core values in regulator facing disclosures.**
    b. **Organizations should stand ready to demonstrate the soundness of the policies and processes they use and how data and data use systems are legal, fair and just.**[17]

[17] Algorithms may be subject to regulatory or certifying body inspection, but would not generally be available to other stakeholders.

Appendix 2

Ethical Decision Making Worksheet for Artificial Intelligence and Machine Learning Environments

The continuing revolution in computing and communications technologies, particularly where observation feeds artificial intelligence, may create pressures on societal values related to equity as well as privacy and data protection. To find balance with multiple interests in this environment the application of ethical frames needs to be captured in business processes. In finding the pathway forward the United Nations 2011 Guiding Principles on Business and Human Rights is helpful. Rather than create a tool kit, the principles created an expectation that businesses should conduct due diligence to protect the rights of others. It requires a business to self-declare how they will avoid causing or contributing to harm to others. This worksheet defines how businesses using artificial intelligence meet the ethical goals of legal, fair and just that link to ethical frames. This worksheet is intended as an aid as businesses cascade their core values through the organisation down to the operational level where ethical objectives will be built into systems.

Hierarchy	Defined	What does it mean in Business terms	What are the interrogation questions? (what a sound organizational process would be evaluated against)	What core or ethical frames does this align to?
Ethic	Cultural norms, values and unspoken beliefs	*Data should serve people, people should expect more than just being the source of digital content. Data and systems should therefore:* • *Maximize the most satisfactory outcomes for the most impacted people, and* • *Maximize social harmony, and* • *Make people the beneficiaries not the targets.* • *The interests of those to whom the data pertain have priority in terms of satisfactory outcomes,*		*Begins argument between consequentialist (focused on outcomes) vs non-consequentialist (focused on intent or duty) ways of making moral judgements.*

		but other stakeholder interests must be considered.		
Values	Translation of values into the core organizational values	• *Our innovation will serve people and society as well as us* • *We will do no harm unless it has societal benefits that may be explained*	• *Does the organization have an "ethics by design process" that is part of the system development process?* • *Are shared organizational values described and/or articulated and can they be reduced to core and guiding principles that are understood by technical staff and that can be programmed into project and data objectives?* • *Have the articulated values been aligned to the varied geo-values across an organizations reach and footprint.*	*Virtue ethics – holds that the goal of life is "ultimate good" and that it is achieved through the application of specific virtues that guide subjective decision making.* • *Are these values consistent across an organization? Or is there room to take into consideration differences across the regions or cultures where an organization does business?* • *Are these values employee based or data source based?*
Core Principles	Expression of values in employee and operational terms	• *Beneficial* • *Fair, Respectful and Just* • *Transparent and Responsive* • *Responsible and Answerable*	• *Does the organization use a comprehensive data impact assessment, a process that looks at the full range of rights and interests of all stakeholders to ensure they achieve a legal, fair and just outcome of data use*? • *Does it assess all risks and benefits to an individual or groups of individuals*? • *Is accountability and responsibility for ensuring legal, fair and just outcomes are achieved established*	• *How are levels of harmony, benefits, and audiences involved determined?* • *How does transparency account for subjective vs objective determinations of values?*

			through clearly defined roles throughout the organization?	
Guiding principles	Expression of values in overarching policies and procedures	***Beneficial*** – • *Uses of data should be proportional in providing benefits and value to individual users of the product or service. While the focus should be on the individual, benefits may also be accrued at a higher level, such as groups of individuals and even society.* • *Where a data use has a potential impact on individuals, the benefit should be defined and assessed against potential risks this use might create.* • *Where data use does not impact individuals, risks, such as adequately protecting the data and reducing the identifiability of an individual, should be identified.* • *Once all risks are identified, appropriate ways to mitigate these risks should be implemented.* ***Fair, Respectful, and Just*** • *The use of data should be viewed by the reasonable individual as consistent, fair and respectful.* • *Data use should support the value of human dignity – that individuals*	• *Where an analytical driven data use or AI design has potential impacts on individuals, are the risks and benefits explicitly defined (significance and likelihood)?* • *Are the risks necessary and proportional to the benefits? Have the risks been mitigated to the extent possible?* • *Are the mitigated risks sufficiently balanced by the benefits?* • *Have the systems themselves, and the data that feeds those systems, been assessed and protected proportional to the risks?* • *Where data processes begin with analytic insights, have those insights been tested for their accuracy, predictability, and consistency with organizational values associated with legal, fair and just principles?* • *Is it foreseeable that the expected insights might seem inappropriate or discriminatory or might be considered offensive causing distress or humiliation?* • *Would individuals be surprised by the systems use of insights about*	*Consequentialism approach (Utilitarian ethics) assesses the likely outcome of an action, policy, or rule and aims to maximize pleasure and minimize pain for the greatest number of people.* • *How are "pleasures" and "pains" translated into "benefits" and "risks"* • *What happens when the achieved outcome does not match the intended outcome?* *Non-consequentialism approach (Deontological ethics) makes judgements based on the intent of an action and aims to preserve equality and human dignity. Guided by 3 questions:* *1) Could your design become international standard or norm?* *2) Does your design* ***provide benefit*** *to individuals involved or does*

		have an innate right to be valued, respected, and to receive ethical treatment. Human dignity goes beyond individual autonomy to interests such as better health and education. • *Entities should assess data and data use against inadvertent, inappropriate bias, or labeling that may have an impact on reputation or the potential to be viewed as discriminatory by individuals.* • *The accuracy and relevancy of data and algorithms used in decision making should be regularly reviewed to reduce errors and uncertainty.* • *Algorithms should be auditable and be monitored and evaluated for discriminatory impacts.* • *Data should be used consistent with the ethical values of the entity.* • *The least data intensive processing should be utilized to effectively meet the data processing objectives.* ***Transparent and Autonomous Protection (engagement and participation)*** • *As part of the dignity value, entities should always take steps to be*	*them? Would they align with the choices they have been provided?* • *Could the design choices become international standards or norms?* • *Does the design benefit individuals and communities or is it designed strictly to benefit from them? For Intensive data impacting systems, does the review assess that outcomes are as predicted, risks are mitigated as planned, harms are reduced, and unintended consequences are understood?* • *Does the review process include a gating by senior accountable leadership? Does this analysis include the likelihood of benefits being achieved and risks effectively mitigated?* • *Does the post review include an assessment of whether or not the anticipated outcomes were achieved?* • *Where internal reviewers need external expertise, is this expertise sought?* • *Have the organizations sought mechanisms that explain how data is used, how the benefits and risks to individuals are associated with the processing, and how individuals may*	*it strictly **extract benefit** from them?* *3) Does the design respect people and treat them the way you would expect to be treated?* *Overarching questions:* *1) Is it appropriate to allow combination of multiple ethical principles into a framework for data processing and what risks are generated by doing so?* *2) Are there universal core societal values? And if there are not how do organizations consider any differences?*

		transparent about their use of data. • *Decisions made and used about an individual should be explainable.* • *Dignity also means providing individuals and users appropriate and meaningful engagement and control over uses of data that impact them.* ***Accountability and Redress Provision*** • *Entities are accountable for their use of data to meet legal requirements and should be accountable for using data consistent with the principles of Beneficial, Fair, Respectful & Just and Transparent & Autonomous Protection. They should stand ready to demonstrate the soundness of their accountability processes to those entities that oversee them.* • *Individuals and users should always have the ability to question the use of data that impacts them and to challenge situations where use is not consistent with the core principles of the entity.* • *They should have accessible and appropriate redress systems available*	*participate and object where appropriate?* • *Have all stakeholder concerns been assessed and appropriately addressed as part of the data system lifecycle.* • *Is the organization ready to demonstrate the soundness of the processes they use so that data and data use systems are legal, fair and just?*	

Rules	Translation of policies into rules	• *Translation of the policies and procedures above into rules and processes*		
Implementation and External Positioning	Rules translated into external commitment to business requirements	• *Translation of the rules into a code of conduct*		