Seminar on Responsible use of Data and AI
Outcome from Working Groups

Introduction to the event and the working groups

DNB, in collaboration with Stanford, PWC, Dataforeningen, DAMA Norge, and BigInsight organized a seminar on Responsible use of Data and AI on January 27th, 2020 in DNB main office, Bjørvika, Oslo. 200 participants represented 37 companies across sectors, industries and roles. The seminar had a morning of eight sessions on relevant topics, a showing of the thought-provoking documentary film iHuman by Tonje Hessen Schei and an afternoon panel debate lead by Torgeir Waterhouse.

After the panel debate, a series of working groups were organized involving major actors in the public/private sector, government, and academia in Norway. The purpose of the working groups was to contribute to establishing a coordinated strategy on responsible use of Data and AI. The number of participants in the working groups varied from six to fourteen (including the facilitator(s)). The participants were well balanced in representing different organizations, industries and roles in each working group, in order to ensure a broad representation of stakeholders to the matters addressed.

The working groups started with a kick-off presentation tailored to frame each of the six working group topics. The discussion was facilitated around several central questions to direct the discussion in a focused approach. As a wrap-up, all the working groups gathered, and each working group facilitator presented the main findings from their respective working groups. The main outcomes from the working groups are summarized below, compiled by DNB. Finally, we present an executive summary including some key cross-sectional outcomes identified across several of the working groups.

The working groups addressed the following topics

Working group 1	Algorithm Governance and Algorithm Audits (For tackling AI misuse)
Working group 2	Strategies and processes for assessment on AI biases and fairness (For tackling unintended consequences from AI)
Working group 3	How to bridge the gaps between technology on the one hand, and law and ethics on the other?
Working group 4	Accountability and explainability under AI (For both Autonomous and Human-Interactive systems)
Working group 5	Ownership, regulation and consent of Data
Working group 6	What can we learn from other fields for AI Regulation?

Summary of the outcome from each working group

The main outcomes from the working groups is summarized below. Disclaimer: The summary may have some inaccuracies that come from interpretation; though it has been endeavored to present the outcomes as correctly as possible. Some editing has been done in an effort to make the notes as understandable as possible for readers who did not participate in the working groups.

Working group 1: Algorithm Governance and Algorithm Audits (For tackling AI misuse)

Warm-up questions

- What governance/audit methods are available to us at different levels?
- What are the pros and cons of the different approaches?
- (How) should the scope of the governance/audit process differ for different applications and/or algorithms?
- Is preventing intentional misuse different from preventing unintentional misuse?
- Who should govern/audit what when it comes to AI systems?
- What should be the role/responsibility of the internal audit function and the other internal lines of defense?
- What should be the role/responsibility of the external auditor and other third-party trust providers?
- What should be the role/responsibility of government agencies?

Discussion points

- **Possible learnings on audits/compliance from safety domains such as DNVGL:**
    - One interesting perspective from companies such as DNVGL is the perception of ML systems. ML are perceived as non-stand-alone but as embedded components of a system. Consequently, risk analyses/audits need to consider the system as a whole, beyond focusing solely on ML algorithms/systems.
    - When performing risk management, DNVGL takes a systemic approach: e.g., what will be the total losses of ship? What are the high-level consequences of failure? e.g., environmental pollution? By defining a system-level analysis, it is possible to define and measure security requirements to avoid systemic consequences.
    - For effective and efficient system-level analyses, DNVGL's strategy is to:
        - Identify critical components, and components that need to be protected
        - Consider different operational scenarios
        - Based on these criteria, prioritize the test cases and explain to the stakeholders why those tests need prioritization

- **Are there any fundamental differences between ML systems audits and other systems that do not use ML? (how much "adjustment" is needed?)**
    - An interesting point is that currently ML is not allowed for critical systems in DNVGL, and current standards are only for manual systems.
    - This implies that for systems involving ML/AI algorithms, new standards need to be defined.
    - Given the variability of rules governing different scenarios, assets and products, the initial impression is that there will be a need for substantial adaptation of current audit methodologies to incorporate ML/AI aspects.
    - Currently DNVGL is slowly trying to integrate ML/AI concerns into system safety assessments in collaboration with their stakeholders/partners, to assess whether to grant certifications on such systems. This is currently a slow, rather complex process.
- **DNVGL example covers safety-related risks, while ML/AI has a wider scope:**
    - It is clear that current applications on ML/AI represent wider risks beyond safety; e.g., criminality, national security, polarization, misinformation, and other adverse consequences to society. These risks need to be outlined by domain-experts in collaboration with ML/AI specialists.
    - The need for domain expertise is illustrated by the current restrictions in the Norwegian criminal system, where it is not possible to use Blackbox results. Data mining is done instead, and then a person who has knowledge of both criminality and statistics would develop, interpret and communicate the results from ML models.
- **Some relevant questions need to be answered before an algorithmic audit can be done. For example:**
    - Who is doing the risk analysis?
    - What kind of risk it is? Reputation risk vs criminal risk?
    - Should be a hierarchy for risks? What should be the grounds for it?
- **An important aspect of ML/AI audits is the explainability of the models. This assumes that is dangerous to use off-the-shelf (or plug-and-play), black-box or automated ML/AI solutions. It is really so? If that is the case, do we need sophisticated explanations or simple feature importance libraries would suffice?**
    - The importance of explainability is highly dependent of the context and potential consequences of decisions that are/not automated by ML/AI.
    - Furthermore, accountability becomes more difficult if the humans in the loop do not understand the why and where the results/predictions come from.
    - Without an understanding of the model and its underlying assumptions, is not clear if the results are accurate.

Summary

- **Agreement/standards for audits are an important first step:** In order to implement effective algorithm audits, we need to define (and agree) upon "what we want to prevent ML algorithms from doing?" (define concrete desired/undesired behavior or purposes).
    - For example, how do you define discrimination?
    - These standards could be defined at general level and could further be specialized according to industry sectors.
- **Learn from existing fields for establishing an audit process:** There is a realm of knowledge and methods from typical risk mitigation approaches (this includes best/worst case scenario analysis). These can be adapted to AI/ML domains.
    - One example is safety regulations in DNVGL.
- **We need to have an open discussion around whether existing regulation is enough or not to prevent AI/ML misuse:** There is a need to have an open discussion around existing regulation (or herein lack of), to prevent AI/ML misuse that can lead to dire consequences. If we are able to determine in what domains/purposes AI should be used (or not used at all), this can be used as basis for discussing potential legislation.
- **How do we get there?** Once different definitions of undesired behavior/usage have been defined, one could use reference data or knowledge- starting with the simplest identifiable case(s) and check if model/system follows (or is compliant) or not? There are already plenty of existing examples of ML/AI usage/implementation that could be analyzed.

Working group 2: Strategies and processes for assessment on AI biases and fairness (For tackling unintended consequences from AI)

Warm-up questions

- How do we agree on a definition of fairness?
- Where do we draw the line between discrimination (not OK?) and indirect discrimination (OK?)?
- How can we detect unfairness?
- Who should detect unfairness?
- What should be done if detected?

Discussion points

- **Example case: COMPAS - correctional offender management profiling for alternative sanctions**
    - A regression model with more than 100 variables for predicting crime.
    - Dressel and Farid: "The accuracy, fairness and limits of predicting recidivism" Science Advances 4.1, 2008.

- COMPAS system has high false positive rate for black defendants.
- Covers different definitions of fairness.

- **Definition of fairness:**
  - Fairness can be defined from a statistically point of view – e.g., accuracy, low False Negative, low False Positive.
  - However, it is also a mathematical term, a feeling and it is political.
  - People on the receiving end feel "unfairness" because they are decided by data, especially by other "similar" people's data.
  - The definition of fairness might differ in a case by case basis, so there is a need to identify the use case a priori. We might also need data to detect the need to assess case by case.
  - The definition of fairness depends on what kind of society we want to build which means that open, public debate might be necessary!
  - The UN's 17 sustainable goals can give a good guideline.
  - Overall, it is a multifaceted concept that requires people with different backgrounds to come together and talk about it.

- **Where do we draw the line between direct discrimination and indirect discrimination?**
  - Direct discrimination: To discriminate on race and gender is not okay.
  - Discrimination depends on what kind of data we put in and what kind of outcomes we are looking for.
  - Some groups of people will "lose" since it is not possible that everyone "wins" – there will be ethical dilemmas where all solutions include suboptimum solutions for some.
  - It is expected that when having enough data, it often alleviates the fairness problem, since it makes it possible to not use discriminative variables.
    - However, how does this balance with other regulations such as "data minimization" maybe there is a need for a hierarchy within regulation.
  - Everyone agrees there is a line, but currently we don't know where it is.

How can we detect unfairness?

- Third party audits - have independent organizations to assess the results.
- Data scientist - making sure your models are representative of the population.
- Detect it statistically – test/verify if your result is correlated with the variables that you should not use.
- Detection of unfairness represents a dilemma - we need to collect more data that to detect it.
- Test the methods with benchmark datasets.
- However, detecting unfairness goes beyond looking at the data.
- Making sure that the data covers the whole population, e.g., do not use regional data to study the whole country.

- **Who should detect unfairness?**
    - Both users and the scientists involved.
    - Journalists.
    - The parties that feel that they are unfairly treated can provide feedback.
- **What should be done if detected?**
    - Is important to have interdependency between both ML and human in-the-loop mechanisms.
    - Establish routines for corrective measures, e.g., model calibration, better data, etc.
    - Continuous improvement after developing and implementing models and methods. This implies that improvement costs need to be reflected in the budget.
    - Establish routines and processes for collecting feedback from the end users. This may also involves educating the staff, so they are aware of unfairness.
    - Investigations and further development of preventive measures.

Summary

- **Public, Open Debate on Fairness:** It is important to have an open, public debate on what is fairness, since this definition will define what kind of society, we will have in the future.

- **Fairness can be Contextual:** The formal definition and identification may be case/context based; therefore, the discussion should also be case driven.

- **Diversity is needed:** The formal definition and identification will require the involvement of diverse backgrounds and expertise areas.

- **Dilemma-solving:** In order to draw the line between direct and indirect discrimination, it is necessary to resolve concrete dilemmas and questions with concrete cases. For that, is necessary to build a knowledge base on known dilemmas around fairness.

- **Feedback loop:** In order to detect unfairness, feedback loop mechanisms need to be established, between those developing the ML/AI solutions and different stakeholders.

- **Measures and processes are needed:** There is a need to establish clear preventive and corrective measures whenever a decision system may represent a risk for unfair treatment.

Working group 3: How to bridge the gaps between technology on the one hand, and law and ethics on the other?

Warm-up questions

- Are there gaps? (e.g., slow implementation of regulation, slow creation of new regulation addressing emerging technology, regulation in accordance to the possibilities/limitations of current technology and the risks involved, difficulties or missing technical solutions for enforcement, or the fact that current tech implementation reflects the moral values of those who develop it?)
- What type of gaps are there? (communication gap, implementation gap, other types of gap?)
- What are the consequences of such gaps?
- If it is determined that there is a gap and the consequences are considerable, what options do we have to close the gap or to handle the consequences?

Discussion points

- **There is a need for assessing the consequences of use of data and AI**
    - Ethical impact assessments can help address and create awareness of possible unethical outcomes. They should be tailored to various sub-disciplines of machine learning and various forms of algorithmic design.
    - Ethical impact assessment can help create awareness and mitigate general negative consequences and should in some circumstances to be conducted at regular intervals (as AI learns and adapts it can change the functioning of products and services over time).
    - Such types of risk assessments can be used to be used to identify, understand and mitigate associated risks, in particular on high risk applications.
- **Diversity and cross functional teams are needed to address the topic of responsible use of data and AI as it is a complex matter with many different types of stakeholders**
    - Establish an ethical council, tailored on the model of The Norwegian Biotechnology Advisory Board. Use it as a forum for gathering different stakeholders to discuss and address ethical questions and implications of use of AI.
    - Representation of the public when addressing the topic is key as they are a critical data source, target for data use/AI and a potentially vulnerable stakeholder with limited awareness of the implications of AI. Further, involvement of the public in anticipatory practices is needed for building trust and legitimacy of AI implementation.
    - There may be suggested a moratorium on particular technology areas where AI is being implemented, such as military autonomous vehicles/robots and particular value-sensitive areas, such as insurance-forecasting on the basis of medical data.
    - Use a sandbox approach to implementation of AI. It can be used to establish enforcement of risks and challenges, the need for regulation, safety risks etc. Can

also enable increased collaboration and better understanding of cross-sectorial issues.
    - Public/Private data sharing needs to be transparent as it is in effect an exchange of assets, and IP on algorithmic training data should be regulated. There is a need for a Code of Conduct on these transactions.
    - 3rd party audits can help collaboration e.g. by increasing transparency but also poses the question of who should have the right to audit and how to ensure they are qualified to do so?
- **There is a significant need for clear and precise definitions**
    - The need for international standards and legislation prerequisites that we find commonly agreed upon definitions of e.g. "AI" and "bias" which are unique and precise yet not so narrow that the area of application is too limited.
    - A definition of AI must be sufficiently flexible to cater for progress in technology while at the same time provide legal certainty.
    - Moreover, training and deliberative practices for better and broader understanding of the societal implications of use of data and AI is needed across the board; for businesses, regulators, government and the public alike as they are all stakeholders in the same issue.
    - Better training and deliberate practices can help create the required knowledge base and awareness of AI which enables us to tackle the question of responsible use.
- **Small data – Norway and similar countries need to take into consideration the implications of datasets being small**

<u>Summary</u>

- **There is a need for defining AI, bias and other central terms**: This will help enable effective regulation, codes of conduct, and discussions. Currently, there are many different definitions which confuses the work on standardizing and safeguarding.

- **The impact of the AI, be it ethical or otherwise must be addressed in (ethical) impact analysis:** When assessing this, there must be a clear separation between assessing the datasets used for training the algorithm and the learning algorithm itself.

- **There is a clear need for collaboration across countries, sectors, industries, roles etc… in order to represent all relevant stakeholders in addressing this topic:** However, that also brings with its complexity, which makes it hard to achieve unique, specific definitions and standards. This is an inherent paradox/Catch 22 that must be navigated.

- **There are several risks associated with the use of data and AI such as:** the loss of freedom, concentration of massive datasets (e.g. the "big 5" in the USA), unclear human liability for the results of AI, who gets to decide what is ethically sound. A specific AI-ethics board could be suggested on a national or supra-national level.

- **Legislation is perceived by several to be too open for interpretation:** There is a need for clearer and more precisely defined legislation on use of data and AI.

Working group 4: Accountability and explainability under AI (For both Autonomous and Human-Interactive systems)

Warm-up questions

- Who takes the blame when AI malfunctions and why it is important?
- What should be the role of explainability techniques when it comes to legal and societal accountability?
- How to explain/justify/argue AI decisions and final decisions?
- How to proof and ensure trust by users and people affected by AI?
- What are possible appeal processes?
- How can appeal processes be scalable?

Discussion points

- **Why do we need explainability?**
    - Trust is a prerequisite for adoption, and explainability can be a way towards building trust in AI.
    - In the other hand, explainability is required by law, and constitutes a human right when people are being handled by companies/organizations using AI.

- **Who do we need to explain our AI to?**
    - Is important to define who are the receivers of the explanation.
        - E.g., Regulators, general public, open to competition?

- **Who should have to explain a model?**
    - The understanding (explanation building) of different parts of an AI/ML system can be assigned to the different actors all along the supply chain.
    - Different entities (should) understand the algorithm, model and data respectively.

- **Accountability under the context of explainability**
    - Maybe there is a need for a third party that can manage accountability?
    - Can we train experts to bridge the gap between model and algorithm?
    - We must define requirements about ethics and trust so that we can require ethics and trust "by design"

- **Example case: Autonomous vessels**
    - Explanations are important for:
        - Machine-Human interactions
        - User comfort
        - Assurance / fail-safe

        - Trustworthiness
    - Case (Autonomous ferry):
        - Passengers wants to know that the ship is aware of hazards.
    - Case (Autonomous cars):
        - How do pedestrians know that an autonomous car has seen them and is acting accordingly?
    - Explanations can also be used to identify bias. Example:
        - Telling doctors from nurses by clothing: OK
        - Telling doctors from nurses by facial features: Probable gender bias
    - Explanations must balance completeness and understandability.
    - We use AI to tackle complexity we don't know how to handle, so it is not always reasonable to demand understandable explanations. The context must be considered carefully in assessing this.
    - Explanations can be more about user experience than about trust, however it is often not sensible to address one separately from the other.

- **Example case: Law enforcement**
    - Law enforcement needs to be agile to adapt to changing rules and regulations.
    - In the other hand, it needs to be very careful about biases, since the available data is often incomplete (mostly data about wrong-doers).
    - Explainability is important when AI is providing decision support.
    - When using AI for decision support, the final decision rests with the case officer, as the officer may have more / other information than the model.
    - Audits are important when dealing with privacy issues.
    - Systems susceptible to feedback loop bias.

- **Challenges in explainability**
    - How do you define an AI model? The government's definition is very broad.
    - What models do we need to explain?
    - Should requirements of explainability depend on potential harm?
        - Example: medical AI vs. Spotify recommendations
    - Is there a difference between the rights to an explanation and the right to an understanding?
    - What is enough explainability?
    - Much of state-of-the-art machine learning is currently too opaque.
    - Much AI is not deep learning and is easier to explain but can still be scary to the public.
    - Dependencies in data sets can make decisions opaque, independent of the algorithms used to reach the decision.
    - How do we explain which correlations are "fair" to take into account in insurance?
    - The model comes from the data, if the data is skewed, the model will be skewed.
    - It is very difficult to correct biases in models without destroying prediction accuracy.
    - If a prediction model works, we are happy, but how do we do robust checks of what works?

        - Analogy to medicine: we often know from experience that something works without knowing how? can we have rigorous testing of correctness that is not based on explanations?

- **Approaches for explainability**
    - Explain parameters and training data.
    - Build trust in the model on trust in the training data and "experience" of the model.
    - Define the stakeholders before defining what needs to be explained.
        - Who should we explain the models to? Why they would want that?
    - Documentation of explanations is a very important aspect; it can be used to perform "due diligence".
    - There should (always) be a human in the loop?
        - Currently one can argue there exist stricter requirements for AI than for humans, for example autonomous vehicles vs. human drivers.
        - We should not aim to automate everything.

Summary

- **Requirements for having a human in the loop:** An open discussion and consensus are needed on what AI tasks should require HIL (human in the loop) examples are:
    - Whenever a decision feeds back into an algorithm, in order to avoid feedback effects and/or amplification/polarization of behaviors/responses.
    - Lifecycle management, e.g. algorithmic audits.

- **Definition of explainability:** An open discussion and consensus are needed on what is explainability.
    - Different groups have different needs; thus, the definition will be different depending of the receiver of the explanation.
        - Customers need a non-technical, intuitive explanation.
        - Developers and Tech workers need to know the drivers.
        - Stakeholders need to have an overview.
    - For processes and systems with human in the loop, explanations are important for all parties involved in the process, not only for the end-receiver.

- **Requirement on explainability:** An open discussion and consensus are needed on what is “sufficient” explainability.
    - This should be contingent of the potential risks and consequences of AI systems.
    - This means that requirements may be in some cases sector specific.
    - For cases such as medical usage, in the absence of sufficient explainability, a rigorous testing framework should be in place.

- **Distributed responsibility for explanation:** The responsibility for explanation lays in technology, and business sides, as well as third parties.

Working group 5: Ownership, regulation and consent of Data

Warm-up questions

- Are there any obstacles in regulation today for desired and appropriate use of AI?
- If yes, what can be done to resolve these obstacles in future legislation?

Discussion points

- **The kick-off presentations concerned the ubiquitous presence of surveillance cameras and how to approach data minimization in a data science use case.**
    - NORCE was used as the contextual setting for discussing camera surveillance, and the regulations governing its use. The addition of the concept of facial recognition technology, and how this technology is regulated and used around the world followed.
    - The second kick-off presentation took form as a presentation of data scientist use case, and how to approach the principle of data minimalization in an artificial intelligence and data analysis context.
- **There is a need for clear definition of "desired and appropriate use"**
- **GDPR and use of data and AI**
    - GDPR is difficult to be in compliance with, and in an AI context especially the principles of purpose limitation and data minimization, but not impossible. GDPR does put some limitations on what AI can be used for, but overall, it's quite flexible.
    - GDPR opens for national legislation as a lawful basis for processing of personal data, and herein lays the key.
    - Processing special categories of data (health data) is not always easy without consent.
- **Other obstacles in legislation**
    - The main obstacles are to be found in the sector specific regulations establishing a statutory duty of confidentiality.
    - There is also a problem when interpreting laws and regulation, where different entities or agencies within the same sector interpret the same law wildly differently.
    - The above combined with duties of confidentiality makes data sharing especially difficult.

Summary

- **Mapping out regulatory needs:** A mapping of the need for establishing national regulation and lawful basis on a sector by sector basis should be undertaken.

- **In the lawmaking process, there should be established a permanent multidisciplinary expert group to assist legislators and law makers** (similar to that

of the Norwegian Biotechnology Advisory Board). This expert group should consist of lawyers with an understanding of technology, technologists who understand what the law needs to regulate and ethicists who understand the broader societal impact of new technologies, including AI.

- **Technology is a tool not the aim:** New laws should be tech-friendly and technology neutral.

- **Confidentiality weighed against common purpose and lawful basis:** Confidentiality should not be an obstacle for data sharing when the purpose and lawful basis for processing are the same for both data controllers.

- **Sandboxes:** The establishment and participation in regulatory sandboxes should be explored.

Working group 6: What can we learn from other fields for AI Regulation? 

Warm-up questions

- Which methods are available which will enable us to ensure that AI development/deployment is safe and responsible?
- If not: Why and how is AI different from other fields?
- Do we need new assurance methods specifically for AI?
- What are we assuring? Algorithms or the context in which an algorithm works?
- Desired outcome: List of tools and methods which can be reused and the limits to these, and reasons for why AI is different from other fields and tools

Discussion points

- **Regulation based on intention vs outcome (e.g. regulation of drugs vs oil and gas) instead of regulating processes.**
    - Is it possible to reuse technology from compliance or cyber security?
    - Opt-in vs opt-out (e.g. organ donors perceive it as too much of a hassle to actively opt in)
    - Lessons learned from health regulation (e.g. health ethics, bio tech): applies citizen panels and ethics is incorporated in some laws.
    - There must be a clear objective behind every regulation: Why do we regulate this field? Is it really needed? Unknown consequences may arise.
- **There is a need for education of both the public and people working with AI about the use of AI.**
    - Many different agencies and institutions should focus on AI legislation: government, constitution, human rights.

  - Many fields such as many types of AI can be hard to regulate, e.g. the car industry and autonomous cars. Who gets to decide the right or ethical actions? Who is liable for the consequences?
  - Culture and competency are important factors in AI regulation; often the most competent people are the most critical.
- **Impact assessments help assess the consequences on different levels.**
  - A Data Privacy Impact Assessment (DPIA) can help considering the risk and impact for the individual.
  - An initial mission analysis can help answer topics as: data minimization, is it within legal framework, and other ethical and legal aspects.
  - A threat or risk assessment is also helpful to try to identify negative consequences upfront.
- **Collaboration is key: how to collaborate with others and learn from their experience? How to build the right partnerships?**
  - Creating best practices and way of working can help standardization of good ways to work with data and AI.
  - Understanding private and public governance and how to engage with regulators is critical.
- **Know your customer (KYC) and anti-money laundering (AML) are two areas of application where data use and AI have great impact but also require carefulness.**
  - The insurance industry typically also applies data in a large extent and has great impact of data and AI (also in potentially harmful or unethical ways).
- **A systems perspective should be applied when considering responsible use of data and AI.**
  - Safety measures should be applied where applicable throughout the system, according to appropriate standards such as ISO.
  - There is a need for better and more extensive standards and regulations of software, as well as methods for technology impact analysis.
  - Corporate accountability needs to be better clarified and management systems regulated.

<u>Summary</u>

- **Regulators: The intention of using data and AI and/or the outcome of it must be regulated – not how it is done.** Product regulation processes (e.g. as applied for food safety) can support this. Moreover, there must be clear objectives behind every regulation, and mindfulness of unknown consequences of legislation. Regulation should be technology neutral an apply a systems perspective. The question arises of which fields should be regulated. Who should be present in fora for legislation on AI? Some identified parties are: The government through the constitution, nations as part of international conventions, companies to define boundaries, and individuals to defend their rights. Agility must be applied for working faster, cheaper, and more effectively.

- **Lessons from the health sector on how to address difficult choices: How and who to decide on what is right is a conundrum also in biotech and health ethics, as well as for regulation of autonomous cars.** Opt-in vs opt-out has proven to have significant effect on the consequences where it is applied, sometimes unforeseen. Ethical citizen councils, in a manner like the Biotechnology Council, can offer a democratic and deliberate approach to assessing these issues.

- **Improvement and awareness: There is a need for educating the public and the people working in the justice system about use of data and AI.** We need to create a culture of responsibility based on a solid knowledge basis. There can be supervised vs unsupervised learning.

- **Partnerships: Where there are shared targets, collaboration is key.** Social responsibility, AML and KYC are areas where this is highly relevant. Through collaboration one should seek to agree upon best practices, ways of work and management culture. Public-private partnerships should be used to support regulatory development. The insurance industry should be involved to identify uninsurable risk.

- **Tools and methods: There are a number of already existing tools and methods which can help in assessing responsible use of data and AI**, such as: DPIA, initial mission analysis, risk assessment (threat assessment), criticality assessment. Moreover, good practices can be achieved by applying risk management practices and tools, methods for technology impact analysis, corporate accountability available tools. To support this, certifications and attestations such as e.g. ISO/ISAE exits and should be expanded upon. There is also a need for improvement of regulation of software products, regulation of data use and data quality standards. Safety requires a combination of methodology and ISO standards. STAMP methods (new method) can be applied. Management system standards should be developed and applied to ensure and proper work processes.

Executive summary with cross-sectional outcomes

In summary, several important outcomes surfaced in the workshops and more than a few were identified across several of the working groups. Below, we summarize some of the most frequently raised conclusions from the discussions.

- **Legislation and regulation:** The discussions underline a large demand for clear, precise and appropriate legislation and regulation with international impact, and interpreted similarly across borders and organizations, entities or agencies. The impact of existing legislation on use of data and AI must be assessed.

- **Standards and ways of work:** In addition to legislation and regulation, there is a need for industry specific and other Codes of Conduct (e.g. for private – public data sharing), sandboxes, ethical councils (taking inspiration from the Norwegian Biotechnology Advisory Board), and other forms of collaboration. Standards must also be developed for audits, where learning from other relevant fields of audits can be applied. Measures and processed should be developed for ethical dilemmas connected to topics such as fairness.

- **Collaborative effort:** Collaboration across nationalities, organizations, roles etc. is key for ensuring proper representation of all relevant stakeholders in addressing responsible use of data and AI. It is necessary to engage in open discussions on e.g. what is fairness in different contexts, and other complex ethical dilemmas. Feedback loops should be implemented to ensure important guidelines and input from different stakeholders to the developers of AI.

- **Definitions:** A prerequisite for standardization and collaboration is achieving unique, precise and shared definitions on AI, bias, explainability, fairness, and other central terms.

- **Impact assessments:** Active and competent use of impact assessments of different types can help mapping out consequences and mitigating negative societal or unethical effects from use of data and AI.

- **Explainability:** Several topics must be discussed with the aim of reaching consensus on with regards to explainability of ML/AI; for example, in which scenarios shall it be mandatory to keep a human in the loop, and what should be the standards and requirements for explainability.