UiO : Faculty of Law
University of Oslo

Artificial Intelligence Risk Management

The risk-based approach in the Artificial Intelligence Act

Candidate number: 631

Submission deadline: 25.04.2022

Number of words: 16 113

Table of contents

1 INTRODUCTION 1

1.1 Thesis topic and research question 1

1.2 Methodology 2

1.3 Thesis structure 4

2 THE ARTIFICIAL INTELLIGENCE ACT 5

2.1 Introduction 5

2.2 Material scope 6

2.3 Territorial scope 7

3 AI – A VERY SHORT INTRODUCTION 7

3.1 Introduction 7

3.2 What is Artificial Intelligence 8

3.2.1 Introduction 8

3.2.2 Machine and Deep Learning 11

3.2.3 Summary 12

3.3 AI – towards a legal definition 13

4 HIGH-RISK SYSTEMS IN THE ARTIFICIAL INTELLIGENCE ACT 14

4.1 Introduction 14

4.2 High-risk systems in Article 6 and Annex III 15

4.3 Amending Annex III 17

4.4 Risks associated with Artificial Intelligence 18

5 THE RISK-BASED APPROACH, AND RISK MANAGEMENT THEORY 20

5.1 Introduction 20

5.2 The risk-based approach 21

5.3 Risk management 24

5.4 Summary 25

6 RISK MANAGEMENT IN THE ARTIFICIAL INTELLIGENCE ACT 26

6.1 Introduction 26

6.2 Risk management system 27

6.3 Data and data governance 29

6.4 Transparency and human oversight 30

7 CASE STUDY FOR AI RISK MANAGEMENT 32

7.1 Introduction 32
7.2 The SyRI judgement 32
7.3 Risks associated with the system 33
7.4 The risk assessment 34
7.5 Risk management measures 36

8 FINAL REMARKS 37

9 TABLE OF REFERENCE 39

9.1 EU-regulation (including proposals) 39
9.2 EU opinions, guidelines, communications, and recommendations 39
9.3 Case Law 41
9.4 Literature 41
9.5 Journals and articles 42
9.6 Online 42
9.7 Audio/Video 42

1 Introduction

1.1 Thesis topic and research question

The topic of this thesis is the relationship between high-risk Artificial Intelligence (AI) systems, as classified in the European Union's forthcoming Artificial Intelligence Act (AIA) articles 6, 7, and Annex III, and risk management theory. The main question is how providers and users of AI systems should conduct a formalised risk assessment, and implement a risk management system in accordance with Chapter 2, AIA – AI risk management (AIRM). The thesis introduces central concepts in risk management theory, and the requirements for risk management in the AIA.

There is no clear answer as to what AI is – how intelligent does a system have to be before deserving the name. However, it is certainly different from the technologies that came before in its complexity, potential, and associated risks. Many European countries have already established ethics oversight bodies to ensure AI development in line with European values and rights.[1] That there is a strong need to regulate AI is beyond doubt.

AI regulation is in its early days. The nature of the technology "requires novel forms of regulatory oversight."[2] As more proposals and recommendations are published, regulators seem to be in favour of the risk-based approach.[3] In short, the risk-based approach to regulation means imposing obligations based on the level of risk. To identify how *risky* something is, be it a single decision, or the development of an entire AI system, requires an AI risk assessment (AIRA). Some regulators, including the EU, are considering implementing AIRA's in their regulatory framework.[4] As of writing, the EU has yet to do so.

AIRA is not a uniform concept. As illustrated in Koene, stakeholders "have proposed different approaches and methodologies for such assessment frameworks."[5] The goal of such frameworks is risk mitigation. The contents of an AIRA framework varies from stakeholder to stakeholder – stakeholders might have different objectives.[6]

The AIA seeks to balance the interests of two parties, namely that of the regulator, and the regulatee. In this case, the regulator is the EU, through the instrument of the AIA. The regulatee, on the other hand, is the legal subject affected by regulation. The AIA effectively requires that risk assessments should be conducted by several parties. These are somewhat different in nature. The regulator will have to make an assessment in order to correctly classify an AI system

[1] Van Roy, et al. (2021) pp. 5

[2] AIA recital 71.

[3] Koene (2021) p. 4.

[4] Ibid. p. 12.

[5] Ibid. p. 5.

[6] Ibid.

as high-risk. The regulatee is subject to requirements and obligations for how to assess and mitigate risks, if a system is indeed considered high-risk.[7] Effectively assessing and mitigating the risks associated with an AI system requires a combination of a legal framework and formalised procedures, such as recommendations laid forth by standards bodies.

Both the risk-based approach and risk management will be explored later in the text. For now, it is sufficient to point out that the risk-based approach is different from the more traditional rights-based approach to regulation. A risk assessment is a technical term within risk management theory. It refers to the process through which risks and contributing factors are identified.

1.2 Methodology

The Artificial Intelligence Act is still a proposal. This thesis is based on the Presidency compromise text, from the Council of the European Union, published on the 29th of November 2021.[8] The text is the first partial compromise proposal submitted by the Council of the European Union (The Council). The weight and value of both proposals and opinions varies. How much weight to attribute various sources, in part, depends on its placement within the EU's ordinary legislative procedure. Article 294 of the Treaty on the Functioning of the European Union (TFEU) details this procedure. In general terms, it consists of rounds of "readings", whereby the Council and the Parliament suggest changes to proposed legal acts, until final adaptation. As of April 2022, the AIA is in its first reading.

The European Commission, as the first step in the ordinary legislative procedure, developed the proposal. The November compromise text was the result of inputs from The Council. The Council, as opposed to the Commission, consists of democratically elected representatives from their respective Member States. The legislative procedure further involves opinions from the relevant Parliamentary committees. The proposed changes to the AIA, expressed in these opinions might indicate what the final text will be. Since the publication of the first proposal, several have issued their opinions. Among these, the Economic and Social Committee (EESC) submitted their first opinion in September 2021.[9] The Committee on Legal Affairs (JURI) submitted a draft opinion[10] to the Committee on Civil Liberties, Justice and Home Affairs (LIBE), and the Committee on the Internal Market and Consumer Protection (IMCO) in early March, in preparation to a joint hearing later that month, addressing their issues, in preparation for the Parliament. Proposed changes vary from committee to committee, meaning that there is, for the time being, no unified text adopted by the European Parliament. The act will likely see update in the near future, as the Parliament considers the opinions of the committees, and eventually adapts

[7] Mahler (2022) p. 254.

[8] Henceforth, the AIA.

[9] See EESC (2021).

[10] See JURI (2021).

the final regulation. However, the Council compromise text from November is, at time of writing, the most complete representation of the AIA. Thus, it is the most suitable text for the work at hand.

With AI regulation still in its early stages, changes will occur, and authoritative sources are scarce – most markedly the lack of case law. As is the nature with proposed regulation, the Court of Justice of the European Union (CJEU) has yet to interpret its provisions.

Throughout the entire legislative process, the Commission sough feedback from stakeholders.[11] With the publication of the proposal, more have uttered their opinion. These do not bear any weight when interpreting the provisions of the AIA, however, inputs from EU bodies such as the European Data Protection Supervisor, and European Data Protection Board, offer value, particularly when highlighting possible weaknesses in the new regulation.

A bulk of this thesis concerns risk management theory. There is no uniform approach to risk management, nor a general agreement of what it entails in practice.[12] The ISO: 31000, and other ISO standards are among the most common tools for risk management. In addition, both the ISO, and several other standards bodies are developing risk management standards specific to AI.[13] This' thesis presentation of risk management theory is mostly based on the ISO standards as presented in Mahler and Gellert. The aim is to construct a clear and suitable model for conducting an illustrative AI risk assessment, in order to understand how it relates, and may be applied to the provisions of the AIA. International standards do not carry any legal weight as such. However, it is clear that in order to fulfil the criteria for a risk management system in the Artificial Intelligence Act, companies will have to adopt some kind of standard, whether those of the ISO, or internally developed frameworks.[14] For this thesis, the choice fell on Mahler and Gellert as primary sources, as they have extensive works on the interplay between risk management and law.

Mahler points to the challenge of combining law and risk management:

> *[M]uch of the dogmatic literature in law can be based on the legal methodology that is relevant for the respective jurisdiction, and there is not always a need to justify the choices about research methodology. […] The challenge with defining a single scientific context for the present study is related to its combination of law on the one hand, and risk management on the other.*[15]

[11] See https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence 'Important Milestones' for illustration.

[12] Mahler (2010) p. 8.

[13] These are mostly in the draft stage, and will not be included in this thesis. See Koene (2021).

[14] See Benjamin, et.al. (2021).

[15] (2010) p. 9.

It is not evident whether the two can be effectively combined. However, I consider the framework of risk management to be the most accessible way to understand the risk-based approach to regulation – how to unpack the AIA, and give an answer to the research question.

Though it will vary depending on the topic, the conclusions drawn in this thesis might be outdated come autumn 2022. General observations regarding AI, the risk-based approach, and risk management will likely stand the test of time. The AIA is a different matter.

The EU already provides for some regulation of AI. Most notable is the General Data Protection Regulation (GDPR) Article 22. Article 35, GDPR further requires impact assessment, similar to the requirements for a risk management system in Article 9, AIA. However, the GDPR will be precluded from this thesis, as the sole focus is the relationship between the AIA and risk management.

1.3 Thesis structure

Section 2 of this thesis briefly introduces the Artificial Intelligence Act, presenting the material and territorial scope of the Act.

Section 3 is a very short introduction to some of the central topics of AI, and AI research. It introduces the definition of AI as it currently stands in the AIA. Secondly, some general concepts. Firstly, the four main approaches to defining AI. Secondly, central techniques in AI development – machine- and deep learning. Lastly, a return to the legal definition of AI in the AIA, including some of the criticism from both Parliamentary Committees, and other stakeholders. This thesis can only scratch the surface of AI research. The presentation is limited to the bare minimum of technological characteristics necessary for an intelligible discussion of high-risk AI systems.

Section 4 regards the main contents of Article 6, 7, and Annex III of the AIA. It aims to give an overview of the systems currently considered high-risk, and what characterises them. As this thesis concerns the regulatee’s obligation to perform risk management, section 4 will not discuss any formal obligation on the part of the regulator – the goal is to provide the necessary context for further discussion on risk management. Finally, in order to illustrate why risk-classification exists, a few examples illustrating risks associated with AI.

Section 5 tackles the risk-based approach to regulation, and risk management theory. The discussion includes the building blocks for how to implement a risk management system, as required by Article 9, AIA. The presentation focuses on the nature of a risk-based approach to regulation, central technical terms in risk management, and methodology for conducting both risk assessment, and the implementation of risk management measures – risk management *sensu stricto*.

Section 6 presents the regulatee's formal obligation to implement a risk management system, as described in Chapter 2, Article 8 and 9, AIA. Further discussion focuses on issues of data governance, and transparency obligations in Articles 10 and 13. It describes how the provisions in Chapter 2 can be viewed as requirements for both risk assessment, and implementation of risk management measures. This thesis precludes the provisions in Chapter 3, AIA.

Section 7 provides a case study for how an AI risk management system, pursuant to Article 9, AIA, might work in practice. It refers to the case of the Dutch *Systeem Risico Indicatie* (SyRI). The Hague District Court deemed the system in violation of Article 8, ECHR, and several provisions of the GDPR.[16] Risks associated with the system are analysed in terms of risk factors and consequences, and possible risk management measures.

2 The Artificial Intelligence Act

2.1 Introduction

April 2021 saw the first publication of the COM(2021) 206 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts' (AIA). Developed by the European Commission, the AIA is part of a larger 'package', which aims to "build a resilient Europe for the Digital Decade."[17] As part of their mission statement, the EU wants to ensure excellence, trust, industrial capacity, and fundamental rights.[18] The 'package' is not without ambition – "new rules and actions to turn Europe into *the* global hub for trustworthy AI."[19] The aim is to "provide Europe with a leading role in setting the global gold standard."[20]

The legal basis for the AIA is the Treaty on the Functioning of the European Union (TFEU) Article 114. It follows from paragraph 1 that the Parliament, and Council, shall adopt measures necessary for the achievement of the objectives in Article 26, and the functioning of the internal market. It falls to the Commission to draw up proposals, in order to achieve this goal.[21] The processing of biometric data has an additional legal basis in Article 16 TFEU – protection of personal data.[22]

[16] Case C-09-550982-HA ZA (SyRI).

[17] https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

[18] Ibid.

[19] Ibid. (my accentuation).

[20] Ibid.

[21] See Articles 114, 26, and 27 in conjunction.

[22] AIA recital 2.

The AIA is a regulation, as defined by TFEU, Article 288. It has “general application […] binding in its entirety and directly applicable in all Member States.” As stated by the Commission in their White Paper on AI, national initiatives lessen legal certainty, and might “prevent the emergence of a dynamic European industry.”[23]

2.2 Material scope

It follows from Article 1, AIA that the regulation provides harmonised rules for placing on the market, and putting into service, AI systems. Furthermore, to prohibit certain systems, and impose requirements and obligations for high-risk systems. As stated in the AIA preamble “artificial intelligence may generate risks and cause harm to public interests and rights that are protected by Union law.”[24] As such, the AIA aims to establish “common normative standards.”[25] The act further identifies certain interests that merit extra attention, namely “health, safety and fundamental rights.”[26] To achieve this goal, the Commission has identified different categories of risk, namely (i) minimal, (ii) limited, (iii) high, and (iv) unacceptable.[27] The AIA explicitly identifies unacceptable (prohibited), and high-risk systems, in Article 5, and Articles 6 and 7, respectively.

Article 2(1) stipulates that the regulation applies to, among others, providers, and users of AI systems. Article 3(2) defines a ‘provider’ as: (i) a natural or legal person, public authority, agency, or other body, (ii) that develops or has developed an AI system, (iii) which it places on the market, (iv) under its own trademark, (v) whether for payment or free of charge. Article 3(4) defines a ‘user’ as: (i) a natural or legal person, public authority, agency, or other body, (ii) using an AI system, (iii) under its own authority.

Important exemptions from the scope of the AIA are systems developed exclusively for military or national security purposes, Article 2(3), AIA. This is in accordance with Article 4(2) the Treaty on European Union (TEU), which leaves national security the responsibility of Member States.[28] Article 2(3), AIA provides further exemption for systems utilised in international law enforcement. Furthermore, the regulation does not apply to systems developed exclusively for

[23] White Paper on AI pp. 2, cf. AIA recital 2. The White Paper, published early 2020, was part of the preparatory works for the eventual AIA proposal.

[24] AIA recital 4.

[25] AIA recital 13.

[26] Ibid.

[27] The number of categories varies depending on source. The AIA explicitly identifies high- and unacceptable risk, whereas lower risk systems are implicit. The official websites add minimal and limited risk.

[28] AIA recital 12.

scientific research and development, cf. Article 2(6), as it should not undermine the freedom of science.[29]

According to Article 52a, AIA, placing onto the market, or putting into service general purpose AI – AI systems “able to perform generally applicable function”[30] does not automatically lead to the system being subject to the provisions of the AIA, unless it has an intended purpose. The reasoning behind this exemption is “to clarify the role of persons who may contribute to the development of AI.”[31]

2.3 Territorial scope

As mentioned, the Artificial Intelligence Act applies to systems placed on the market, or put into service ‘in the Union’, cf. Article 1(a). Article 2 further specifies that the regulation applies to providers irrespective of whether present in the Union or a third country, if the system is placed on the market, or put into service in the Union, cf. Article 2(1)(a), or where the output produced by the system is used in the Union, cf. Article 2(1)(c). The regulation only applies to users if present, or operating in the Union, cf. Article 2(1)(b).

Article 114 TFEU does not have a corresponding provision in the Agreement on the European Economic Area (EEA). As a starting point, it is therefore clear that the AIA is not binding for EEA countries. That does not deny the possibility of harmonising, or implementation of similar regulation. Among EEA members “On 10 April 2018, 24 [EU] Member States and Norway committed to working together on AI.”[32] Whether or not the AIA applies directly, it is safe to assume that many, if not most AI systems developed within the EEA will be subject to the AIA, cf. Article 2(1).

3 AI – a very short introduction

3.1 Introduction

As stated in the AIA preamble, “[t]he notion of AI systems should be clearly defined to ensure legal certainty, while providing the flexibility to accommodate future technological developments.”[33] The legal definition of AI should be such as to “distinguish it from more classic software systems and programming.”[34] The AIA provides a *possible* legal definition of AI in Article

[29] AIA recital 12a.

[30] AIA recital 70a.

[31] Ibid.

[32] AI for Europe pp. 2.

[33] AIA recital 6.

[34] Ibid.

3(1). The provision was subject to major revision – entirely re-formulated, from April to November 2021.

According to the Article 3(1)(i) an AI system receives machine and/or human-based data inputs. Article 3(32) provides a circular definition of 'input data' as "data provided to or directly acquired by an AI system on the basis of which the system produces an output." Secondly, according to Article 3(1)(ii), AI infers how to achieve a given set of human-defined objectives using learning, reasoning or modelling implemented with the techniques and approaches listed in Annex I. Thirdly, AI generates outputs in the form of content, predictions, recommendations or decisions, which influence the environments with which it interacts, Article 3(1)(iii).

The techniques listed in Annex I include, (i) machine learning, (ii) reinforcement learning, (iii) logic- and knowledge-based approaches, (iv) statistical approaches, (v) Bayesian estimation, and (vi) search and optimization.

According to Article 4, the Commission is empowered to amend the list in Annex I. Addition of techniques requires that they are within the scope of the definition in Article 3(1), and that they are similar to the ones already included in the Annex.

Data, inputs and outputs, machine and reinforcement learning – none of these are legal terms, nor can they be explained through legal reasoning. Yet, a general understanding of these terms and concepts are crucial to see the material reach of the AIA, risks associated with AI, and why AI should be regulated. This section begins with a brief introduction to the definition of AI. Secondly, key techniques for AI development, namely machine- and deep learning. Lastly, an evaluation of the proposed legal definition in Article 3(1) AIA.

3.2 What is Artificial Intelligence

3.2.1 Introduction

In 1901, whilst exploring a shipwreck of the coast of Antikythera, divers discovered a piece of eroded bronze. Later analysis revealed an intricate mechanism, dating from between 200-80 BC. By turning a crank, the mechanism would set gears in motion, accurately predicting the position of astronomical bodies, eclipses, and athletic games, decades in advance. Providing an 'input' resulted in the computation of an 'output' – a computer. Around the same time, approximately 200 BC, Ktesibios of Alexandria built a self-regulating water clock, which maintained a constant flow. Probably the earliest example of an artefact adapting to its environment.[35] Even with the development of machines that run on electricity, microchips, and advanced software,

[35] Russell (2022) p. 33.

this basic formula of input, computation, and output is the same today as it was more than 2000 years ago. What separates modern computing, and by extension AI, from these artefacts is the amount of data (input), the complexity of computation, possible results (output), and possible ways to adapt.

Computers are beginning to surpass the human brain:

	Supercomputer	Personal Computer	Human Brain
Computational Units	10^6 GPUs + CPUs	8 CPU cores	10^6 columns
	10^{15} transistors	10^{10} transistors	10^{11} neurons
Storage units	10^{16} bytes RAM	10^{10} bytes RAM	10^{11} neurons
	10^{17} bytes disk	10^{12} bytes disk	10^{14} synapses
Cycle time	10^{-9} sec	10^{-9} sec	10^{-3} sec
Operations/sec	10^{18}	10^{10}	10^{17}

Figure 1 – man vs. machine[36]

As the table shows, the human brain is still superior to the personal computer in most metrics. On the other hand, state of the art supercomputers have surpassed the human brain in both storage and processing speed. Amount of storage determines how much input the computer/brain can handle. Computational units, cycle time, and operations per second determine the speed and complexity of computations – outputs by extension. This awesome power is not only necessary for its development, but also indicative of how AI systems can baffle, and seem overly complex.

In general, AI "refers to systems that display *intelligent behaviour* by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals."[37] It is possible to distinguish between two main types: weak, and strong AI. Weak AI refers to machines that act in a seemingly intelligent way. Strong AI, later dubbed human-like, or general AI, are machines that are actually conscious.[38] There are several possible approaches and corresponding definitions of AI. Each approach is focused on different understandings of intelligence. We will briefly address four of these approaches.

3.2.1.1 Acting humanly – 'the Turing test'

Devised by Alan Turing in 1950, the test is an early formulation, or method of testing AI. The test seeks to answer the question of whether a machine can think. It consists of an interrogation, where the goal is for the machine to fool the interrogator into believing that it is in fact human.

[36] Ibid. p. 31.

[37] AI for Europe p. 1 (my emphasis).

[38] Russell (2022) p. 1032. Note that Article 52a, AIA refers to *General purpose AI* – a completely different concept.

Passing the Turing requires at least the following capabilities: (i) natural language processing (human communication), (ii) knowledge representation (storing knowledge – memory), (iii) automated reasoning (answer questions, and draw conclusions), (iv) machine learning (adapt, and recognize patterns).[39] Additionally, the 'total Turing test' requires physical human interaction, which requires (i) vision, (ii) robotics. As pointed out in Russel and Norvig, the Turing test has not been subject to much study, researchers rather opting for the practical development of AI.[40]

3.2.1.2 Thinking humanly – cognitive modelling

In order to determine whether a machine "thinks like a human, we must know how humans think."[41] Identifying AI through cognitive modelling requires extensive research into 'cognition'. This can be achieved through observation of human thought, 'introspection', psychological experiments, and brain imaging. This approach combines AI research with cognitive science.[42]

3.2.1.3 Rational thought

Rational thought is to some extent interchangeable with logic – that is "precise notation for statements about objects in the world and the relation among them."[43] In terms of AI, it means a program that can solve any problem that lends itself to logical notation. This is achieved through the combination of accurate knowledge of the world, and probability – making inferences of results based on assumptions.[44]

3.2.1.4 Acting rationally

Acting rationally requires a rational 'agent'. In computer terms, the rational agent is a program that can "operate autonomously, perceive their environment, persist over a prolonged time period, adapt to change, and create and pursue goals."[45] The rational agent seeks the best outcome – in the case of uncertainty, the best expected outcome.[46] The rational agent approach is more practical than the others, as rationality can be achieved in several ways. Correct inference, through rational thought is one of many possibilities. 'Rationality' can also be clearly defined through mathematical formulae or defined goals.[47] The application of statistical models, probability theory, and machine learning enables the program to reach the best expected outcome,

[39] Ibid. p. 20

[40] Ibid.

[41] Ibid.

[42] Ibid. pp. 20-21.

[43] Ibid. p. 21.

[44] Ibid.

[45] Ibid. pp. 21-22.

[46] Ibid. p. 22.

[47] Ibid.

based on available data. This is the pervasive approach to AI, "*the study and construction of agents that* ***do the right thing***."[48] It has therefore been dubbed the 'standard model' of AI. The AIA, both in its definition of, and regulation of AI, is clearly focused on the regulation of AI systems that fall under the 'standard model'. However, this focus might be unfortunate, as "the standard model assumes that we will supply a fully specified objective to the machine."[49]

3.2.2 Machine and Deep Learning

In its most simplistic formulation, machine learning is an automated statistical analysis of correlation and probability, leading to conclusions about the world. It means that the agent/machine develops and improves, as it accrues more knowledge.[50] There are three main types of learning. 'Supervised learning' means supplying the machine with an input-output pair. By labelling the outputs, the machine learns the characteristics of an input that would lead to said output.[51] In turn, it is possible to learn the correct output for an unlabelled input. Feed the machine 50, labelled images of a cat, followed by an un-labelled umbrella. The machine concludes that an umbrella is not a cat. 'Unsupervised learning' means that the data set is unlabelled, nor does the machine receive any feedback – affirmation that it has reached the right conclusion. This type of learning is commonly used for 'clustering' data – finding correlations.[52] However, the machine will not necessarily find the correct causation. We will take a slightly more detailed look at the third type of learning, reinforcement learning.

There are a possible 255,168 games of Tic Tac Toe (Noughts and Crosses), yet, an adult may easily conclude that, bar any amateurish moves, all games end in a tie. If 'traditional' software was programmed to play the game, it is likely that the designer would feed the program with not just the rules of the game, but also general recommendations of which move to perform in light of the current position. Allowing the machine to learn by itself however, functions very differently. With deep learning applied to the game of tic-tac-toe, the software would simply get a pat on the back if it were to win the game – 'reinforcement learning'. It would not even need to know the rules of the game. Through 'inverse reinforcement learning', the machine is able to learn the rules (discover the utility function) by simply observing players. The software would then go on to play large numbers of games against itself. Analysis of the resulting games would then lead to the conclusion of how best to play. Inverse reinforcement was applied in the development of AlphaZero, the world's most powerful chess computer. Later, the development

[48] Ibid. p. 22 original emphasis.

[49] Ibid. The question of how to correctly specify an objective can have major implications for the functioning of an AI system. See the 'value alignment problem' later in the text.

[50] Ibid. p. 668.

[51] Ibid. p. 671.

[52] Ibid.

of AlphaGo, and subsequent defeat of world-class Go players.[53] Deep learning algorithms have defeated human players in a variety of complex games, such as DotA 2, a game with literally endless amounts of possible ‘moves’ and interactions, and Jeopardy!, which required the machine not just to find the answer to a question, but according to the rules of the game, infer the correct question.

Deep learning has its roots in early attempts to model the firing of neurons. Therefore, it is sometimes referred to as neural networks.[54] “**Deep learning** is a broad family of techniques for machine learning in which hypotheses take the form of complex algebraic circuits with tunable connection strengths.”[55] It means long computation paths – outputs are fed back in to the inputs –a recurrent network.[56] The opposite of a recurrent- is a feedforward network. The figure below illustrates the difference between feed forward- (a, and b) and a recurrent network (c).

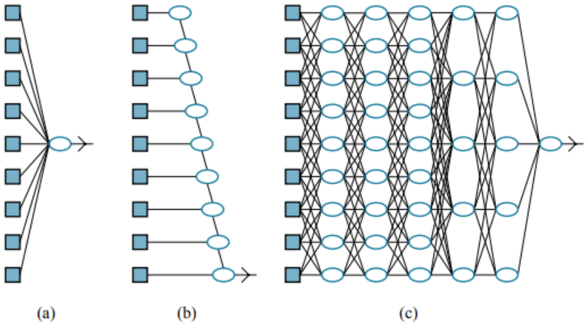

Figure 2 – Feedforward vs. Recurrent Networks[57]

For the purposes of this thesis, it is important to note that the computational paths of recurrent networks, deep learning, are more difficult to interpret and explain than those of feedforward networks. It leads to questions of lacking transparency, or makes it difficult to explain why the machine reached a certain decision, i.e. whether to grant a loan.

3.2.3 Summary

In order to determine whether machines are actually displaying intelligent behaviour, we can compare the performance of AI to humans. AI now outperforms humans in ‘object detection’ (correctly identifying objects in a picture). High-end AI has surpassed human ability to correctly answer questions. Humans have suffered a series of defeats to AI in games such as chess and

[53] In 1950, Claude E. Shannon calculated that there are around 100^{120} possible games of chess. By comparison, the observable Universe contains approximately 10^{82} atoms (Baker 2021). The game of Go has even more possible games, making the achievement of AlphaGo even more impressive.

[54] Russell (2022) p. 801.

[55] Ibid.

[56] Ibid. p. 802.

[57] Ibid.

poker, but also more advanced games, with even more possible actions and outcomes, i.e. Jeopardy!, Go, DotA 2, and StarCraft II.[58] Algorithms are complex, computational paths of recurrent networks difficult to interpret; there is no guarantee that humans will understand the workings of the machines of the future. Herbert Simon summarises this perfectly:

> *It is not my aim to surprise or shock you – but the simplest way I can summarize is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the human mind has been applied.*[59]

3.3 AI – towards a legal definition

The European Economic and Social Committee recommended a clarification of Article 3(1), AIA, as well as removing Annex I in its entirety.[60] As they point out in their opinion, "a number of the examples given in Annex I are not considered AI by AI scientists, and a number of important AI techniques are missing."[61] The Committee on Legal Affairs argue for the removal of Annex I, points (b) and (c), whilst amending point (a). Reasoning that the justification for the AIA was the regulation of "(rather new) machine-learning and data-driven AI applications."[62]

According to the definition in Article 3(1)(i), AIA, an AI system receives inputs. On the surface, this seems superfluous. However, I would argue that it represents a limitation in the definition of AI. It is not at all obvious that the system would need inputs/data per-se, other than during original programming. The reference to Annex I is, as stated by the EESC, a major error. Firstly, they rightly point out that not all the techniques, e.g., statistical approaches, are AI in and of themselves. Nor is it necessary for an AI system to generate outputs which influence the environments it interacts with – AI is characterised by the system itself, not what it produces.

The Parliament has asked for the inclusion of systems produced by AI.[63] It is unclear whether such systems actually fall outside the scope of the AIA. As mentioned, both providers and users are defined as a natural or legal person, public authority, agency, or other body. The wording would suggest no, granted that an AI system capable of producing a new system has 'provided'

[58] Russell (2022) p. 45-46.

[59] Herbert Simon (1957), in Russell (2022) p. 39.

[60] EESC (2021) p. 3.

[61] Ibid. p. 6.

[62] JURI (2021) p. 139.

[63] 2020/2012(INL) pt. 6.

said system. On the other hand, the new system would also, in a technical sense, be an output. Risks associated with outputs lie at the very heart of the AIA.[64]

The overall goal of the AIA is to regulate advanced technologies, that have major implications in the present and the future. In order to achieve this goal, it is not necessary to 'lock' the AIA to a strict definition of AI. This would run the risk of omitting systems that have similar effects, without displaying intelligent behaviour. The goal, seen in conjunction with the necessity for a clear definition of AI, leads me to the conclusion that the AIA should rather be re-named The Advanced Software and Artificial Intelligence Act.[65]

4 High-risk systems in the Artificial Intelligence Act

4.1 Introduction

Among its aims, the AIA seeks to protect public interest and fundamental rights.[66] Whilst the EU maintains that "most AI systems pose limited to no risk […] certain AI systems create risks that need to be addressed to avoid undesirable outcomes."[67] With the potential risks posed by AI, the European Council prompted the Commission to "provide a clear, objective definition of high-risk [AI]."[68] The solution presented is "a clearly defined risk-based approach […which should] tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate."[69] The risk-based approach envisioned in the AIA operates with four risk-categories, namely: minimal, limited, high, and unacceptable risk. Depending on the category, the AIA imposes requirements, or outright prohibition. The first two categories, minimal, and limited risk are, with the exception of limited transparency obligations, subject to few requirements.[70] Requirements for the third category, high-risk systems, compose the majority of the AIA. The fourth category, unacceptable, is subject to outright ban.[71] The Commission's proposal considers this to be in accordance with the principle of proportionality, as the risk-based approach only "imposes regulatory burdens […] when an AI system is likely to pose high risk to fundamental rights and safety."[72] Mahler questions whether the AIA is in fact aimed at

[64] Though not necessarily the definition of AI.

[65] Similar position as Schwemer, see Jon Bing Memorial Seminar.

[66] AIA recital 13.

[67] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[68] EUCO 13/20 p. 6.

[69] AIA recital 14.

[70] See Article 52, AIA.

[71] AIA recital 14.

[72] AIA Memorandum 2.2.

managing risks associated with AI, or if the motivation behind the regulation is strictly to limit the regulatory burden of AI providers – a way to avoid a strict blanket regulation.[73]

We can distinguish between two types of AI risk assessments. Firstly, a traditional risk assessment. Secondly, risk classification. Risk classification, as illustrated by the AIA, "is specifically relevant in the context of law."[74] Risk classification is an abstract form of risk management. It means that risks are not managed per se, but rather that the classification prompts the regulatee to conduct a formal risk assessment.[75] In accordance with the risk-based approach envisioned in the AIA, risk classification is the instrument through which the regulator imposes legal obligations. As such, an overview of risk classification, in this case limited to high-risk systems, is part of the basis for further discussion on the obligation to conduct a formal risk assessment, and implement risk management procedures.

The following sections introduces Articles 6, 7, and Annex III. It aims to give an overview of the criteria for high-risk classification, and the systems currently included in Annex III. Furthermore, it includes examples of risks associated with AI – examples that illustrate the rationale for the classification of some systems as high-risk.

4.2 High-risk systems in Article 6 and Annex III

Article 6, AIA, lists the classification rules for high-risk AI systems. Article 6(1) and (2) regard product safety and harmonisation legislation.[76] Article 6(3) stipulates that AI systems included in Annex III "shall be considered high-risk."

Early in the development of the AIA, in their White Paper, the Commission stressed that there should be "clear criteria to differentiate between different AI applications."[77] The riskiness of an AI system should be determined by "what is at stake, considering both the sector <u>and</u> the intended use"[78] For reasons of legal certainty, the Commission wished for an exhaustive list of AI systems – those deemed risky enough.[79] The AIA provides this exhaustive list in Annex III.

The list includes eight areas, where the use of AI can involve a high risk. These are (i) biometrics, (ii) critical infrastructure, (iii) education and vocational training, (iv) employment, workers management and access to self-employment, (v) access to and enjoyment of private services

[73] Mahler (2022) pp. 248-249.

[74] Koene (2021) p. 5.

[75] Mahler (2022) pp. 254-255.

[76] Which falls outside the scope of this thesis.

[77] White Paper on AI pp. 17.

[78] Ibid. (original emphasis).

[79] Ibid.

and public services and benefits, (vi) law enforcement, (vii) migration, asylum and border control management, (viii) administration of justice and democratic processes.

Among their common denominators are a high risk to fundamental rights, and risk of harm to health and safety. Systems that pose a high risk in terms of severity and probability. According to the AIRA preamble, the Commission identified these eight areas using the same methodology as envisioned in Article 7.[80] This implies that the formulation of Annex III included the Commission conducting several, or at the very least eight, risk assessments. How the Commission decided on the eight areas is rather unclear. For instance, Mahler questions whether an actual assessment took place.[81]

Risks associated with each of the areas in Annex III are mentioned in the preamble, recitals 33-40, AIA. To illustrate, "failure or malfunctioning [of systems managing infrastructure] may put at risk the life and health of persons at large scale."[82] Systems that determine access to education may, "improperly designed and used […] violate the right to education […] and perpetuate historical patterns of discrimination."[83] Improperly trained, systems applied to law enforcement, as shown later in this thesis, can "single out people in a discriminatory […] manner."[84] Systems used within administration of justice and democratic processes represent a high-risk "considering their potentially significant impact on democracy, rule of law [etc.]"[85]

The eight areas are wide in their scope. Feedback from Parliamentary Committees indicate that Article 6, and Annex III will undergo major changes before the European Parliament passes a final regulation. The EESC has been most critical of what they refer to as the "list-based" approach in Annex III, stating that: "it runs the risk of normalising and mainstreaming a number of AI systems and uses that are still heavily criticised."[86] Indicating a wish to remove Annex III. In their draft opinion, The Committee on Legal Affairs (JURI) argue that Article 6, and Annex III should be in line with the approach in the Commissions White paper, where high-risk classification was not only dependant on the sector, but also where the intended use involved a *significant* risk.[87] Because Annex III is "way too broad and vague" JURI forcefully state, "AI systems with hardly any risk would face the burdensome obligations of Chapter 2. In other words, this is not a risk-based approach!" hardly in line with "[a]n EU that strives for

[80] AIA recital 32. See below for Article 7.

[81] See Mahler (2022) pp. 264-.

[82] AIA recital 34.

[83] AIA recital 35.

[84] AIA recital 38.

[85] AIA recital 40.

[86] EESC (2021) p. 4.

[87] JURI p. 41, cf. White Paper on AI p. 17.

global leadership in AI" with "the right balance between promoting innovation and protecting fundamental rights."[88]

4.3 Amending Annex III

Article 7, AIA contains the provisions for amending Annex III. According to Article 7(1), the Commission may adopt delegated acts, in accordance with Article 73 to update the list in Annex III.[89] An update, or rather an addition of a high-risk system to Annex III is subject to two cumulative criteria. Firstly, as stipulated by Article 7(1)(a), the system's intended use must be within one of the areas listed in Annex III, one through eight. Secondly, it follows from 7(1)(b), that the system must, (i) pose a risk of harm to health and safety, or (ii) a risk of adverse impact to fundamental rights, which lastly (iii) must be equivalent to, or greater than the risk posed by systems already included in Annex III.

Article 7(2)(a)-(h) un-exhaustively lists criteria for determining whether a system poses a risk to health and safety, or an adverse impact to fundamental rights. A full presentation of all these criteria fall beyond the scope of this thesis. I therefore limit the analysis to the notion of 'adverse impact to fundamental rights'.

An adverse impact on fundamental rights is the second condition under which a system may receive high-risk classification under Article 7, AIA. It is unclear how the Commission should determine whether an AI system has an 'adverse' impact. If something is 'adverse', it means that it is harmful, or unfavourable. This is modified by the criteria that the 'adverse' impact has to be equal to, or greater than those systems already included in Annex III. As pointed out in the previous section, parliamentary committees found Annex III gravely lacking. However, the preamble notes that high-risk classification should be limited to systems that "have a significant harmful impact."[90] This indicates that there is some threshold to be crossed before the criteria of 'adversity' is fulfilled.

The preamble points to "the protection of fundamental rights, as recognised and protected by Union law."[91] Maintaining fundamental rights are among the founding principles of the European Union. As stipulated by the Treaty on European Union (TEU) article F(2), "The Union shall respect fundamental rights, as guaranteed by the [European Convention on Human

[88] Ibid. (Original in *italic*).

[89] The proposal has received further criticism for delegating too much power to the Commission. Cf. structure of the Artificial Intelligence Board, Articles 56-58, AIA. See i.e. JURI (2021) p. 43 with suggested amendment to Article 7(1), AIA, and EDPB-EDPS (2021) pp. 15-17.

[90] AIA recital 27.

[91] AIA recital 5.

Rights].” Furthermore, the normative standards that the AIA seeks to implement should be consistent with the Charter of fundamental rights of the European Union (The Charter).[92] The Charter seeks to reaffirm and strengthen the fundamental rights afforded to citizens of the Union.[93] Rights enshrined in the Charter stemming from the ECHR have the same scope and meaning.

What ‘adverse’ means and how ‘adverse’ an AI system has to be before being considered high-risk are questions that will undoubtedly be answered before Parliament passes the final AIA. As it currently stands, a plausible conclusion is that the Commission will have to apply the same methodology as the CJEU and the European Court of Human Right (ECtHR) in their interpretation of the Charter and the ECHR respectively. Whether a high-risk system has an adverse impact on fundamental rights would then mean a test of (i) legal grounds, (ii) whether necessary in a democratic society, (iii) a legitimate aim, and (iv) proportionality.

In their joint opinion, the European Data Protection Board and the European Data Protection Supervisor note that “collective effects” should also be included.[94] According to the EDPB-EDPS, some provisions of the AIA is lacking in its protection of “groups of individuals or the society as a whole.” Pointing to group discrimination, and public discourse.[95]

The purpose of risk-classification is a system whereby risks can be mitigated. Before moving on to the ways in which risks are mitigated – risk management, I will discuss a very select few of the risks currently associated with Artificial Intelligence.

4.4 Risks associated with Artificial Intelligence

There are many examples available to identify some of the unique risks associated with AI systems. Documented cases include discrimination, privacy intrusion, and systems that have tremendous derogatory effects, such as social scoring.[96] Of the negative effects of AI seen thus far, many originate in either design, or the choice of data.[97] Further development, and new applications will reveal hitherto unforeseen risks, thus “the individual and societal effects of AI systems are, to a large extent, unexperienced.”[98]

One of the central risks associated with AI is the ‘value alignment problem’, dubbed the ‘King Midas problem’. As was the case with King Midas, the question of value alignment can be

[92] AIA recital 13 (see also Memorandum 1.2). The Charter only applies to Members of the Union (EEA/EFTA excluded).

[93] AIA recital 4 and 5.

[94] EDPB-EDPS (2021) p. 9.

[95] Ibid. Article 7(2)(d), AIA does, however mention “plurality of persons”.

[96] Koene (2021) p. 8.

[97] White Paper on AI p. 10.

[98] EDPB-EPDS (2021) p. 6.

boiled down to: *be careful what you wish for*.[99] This problem arises "when a utility function fails to capture background societal norms."[100] Consider an AI system designed to play a video game. Documented cases involve AI avoiding defeat by purposefully crashing the game. When the rules were modified to prevent such behaviour, the AI figured out how to crash the opponent's game instead.[101] In each case, the objective was to win the game; the designers had neglected to adequately specify all acceptable ways to victory, however. Exemplified as a research exercise, with AI trained to play video games, 'value alignment' sounds like an insignificant problem. However, all the possible discrepancies between what the developer wants, and what the developer gets, lies at the heart of AI risk.

Section 3 illustrated the complexity, and sometimes difficulty of interpreting output. In a legal context, it may erode "our capability to give a casual interpretation to outcomes, in such a way that the notions of transparency, human control, accountability and liability over results will be severely challenged."[102] Interpreting a simple if/then model, such as a decision tree, does not pose much of a challenge. Trying to interpret how a neural network/deep learning produced a certain output is an entirely different matter. Interpreting the output of an image recognition program might result in, "after processing the convolutional layers, the activation for the *dog* output in the softmax layer was higher than any other class."[103] Artificial neural networks simply do not work in the same way as human ones, nor can they be held to account.[104] In order to fix this problem, an adequate level of 'explainability' might require a separate program, able to interpret the processes of the first, producing a less esoteric answer to the question of why.[105] I.e. the animal had fur and four legs. An alternative is 'explainable AI' (XAI), systems able to explain their own processes in intelligible language.[106]

Human oversight and accountability directly relate to transparency and explainability. The challenge of human oversight over complex systems is self-evident. In relation to this, AI brings with it the very real fear of automation bias – the user putting too much trust in the machine. The EESC argues that the AIA opens the gate for fully automated decisions, provided that systems adhere to set requirements.[107] To avoid the problem of automation bias, the EESC suggests that some decisions should remain fully in human hands, "particularly in domains where these decisions have a moral component and legal implications or a societal impact such as in the

[99] Russell (2022) p. 1054.

[100] Ibid.

[101] Krakovna (2018) in Russell (2022) p. 1054.

[102] EDPB-EPDS (2021) p. 6.

[103] Russell (2022) p. 729.

[104] EDPB-EDPS (2021) p. 6.

[105] Russell (2022) pp. 729-730.

[106] Ibid. p. 1048.

[107] (2021) p. 3.

judiciary, law enforcement, social services, healthcare, housing, financial services, labour relations and education."[108]

Among the risks associated with AI, is the possibility that algorithms might perpetuate societal bias – algorithmic discrimination. The fear of algorithmic discrimination does not stem from bad design or malicious intent, it is a question of which correlations the program makes when faced with a large dataset.[109]

Algorithms have made their way into law enforcement, through 'predictive policing', sentencing, setting bail, and several other areas.[110] By analysing data, such as correlation between previous behaviour and recidivism, algorithms predict future behaviour. This, in turn can either lead to, or automate a decision regarding, i.e., bail.[111] Algorithms sometimes base their decisions on the *wrong* data. The dataset supplied might include categories such as race, sex, or age, which should not be taken into account. Yet, if not supervised, an algorithm designed for such a purpose might determine such factors to be relevant. Furthermore, the data only includes those who have actually been sentenced. If one group has historically received more lenient sentences, or are acquitted at higher rates, other groups will be overrepresented.[112] For example, an algorithm used by the US justice system led to great discrepancies between black and white defendants. Black defendants were twice as likely to be labelled as repeat offenders, and given a higher risk-score.[113]

5 The risk-based approach, and risk management theory

5.1 Introduction

Adapting a risk-based approach is synonymous with the application of risk management. In many fields, risk management is well established and refined. Risk management in law however, is still a relatively novel concept.[114] Whilst terms such as risk analysis and management are explained later in the text, a general understanding of risk is useful at this point. Mahler points to varying definitions of risk, as they appear in different disciplines. Easiest to grasp is the technical, *realist* position to risk. It generally assumes that risk is a concrete, quantifiable concept, which lends itself to empirical study, where results can be refined with additional

[108] (2021) p. 3.

[109] White Paper on AI p. 12.

[110] Hannah-Moffat (2019) p. 454.

[111] Hannah-Moffat (2019) p. 459.

[112] Russell (2022) p. 1045.

[113] Angwin (2016). The developers vehemently disagreed that this was the case.

[114] Mahler (2010) p. 1.

data.[115] Secondly, the economic perspective is recognizable through its distinction between risk and probability.[116] Probability means the inclusion of assumptions, in this case probable outcomes a priori, as part of a data set, whilst correcting as more data becomes available. A legal risk is “a risk that has a legal issue as its source.”[117] For instance, litigation, damages, or non-compliance.

In terms of law, risk is already a term filled with meaning. She “assumed the risk”, the seller “bears the risk” whilst the goods are in transit, he “put her at risk”. As the risk-based approach, and risk management theory creeps into the field of law, so too does its nomenclature – though not always consistently. The Artificial Intelligence Act mentions ‘risk’ 344 times.[118] However, a clear definition, or consistent use, is sorely lacking.[119]

Among the most widely used tools for risk management are the ISO standards. They are not only employed by companies, but also influence statements and guidelines from public authorities. The European Data Protection Board’s, The Guidelines on Data Protection Impact Assessment (DPIA) follow the ISO standards, and gives this *standard* definition of risk, and risk management.

> *A “risk” is a scenario describing an event and its consequences, estimated in terms of severity and likelihood. “Risk management”, on the other hand, can be defined as the coordinated activities to direct and control an organization with regard to risk.*[120]

5.2 The risk-based approach

A *traditional* conception of the field of law is the regulation of rights – something is legal or illegal, in computer terms: true or false. As Gellert illustrates, the risk-based approach is fundamentally different. Rather than binary, the risk-based approach is scalable, the goal: to identify an (un)acceptable level of risk.[121] Furthermore, the rights-based approach does not normally distinguish between different legal subjects, thus “the same rules [apply] to everyone.”[122] The risk-based approach however, imposes legal obligations depending on the *level* of risk. It “involves the use of a systematized framework of risk classification to categorize type and degree

[115] Mahler (2010) p. 30.

[116] Ibid. p. 31.

[117] Ibid. p. 84. For a comprehensive discussion of different definitions, see Mahler chapter 6.

[118] Mahler (2022) p. 248.

[119] See ibid. pp. 259-263.

[120] Art 29 WP (2017) p. 6. Guidelines for impact assessment pursuant to GDPR Article 35.

[121] (2020) p. 2.

[122] Gellert (2020) p. 2.

of risk posed by the object or activity being regulated."[123] Gellert (on data-protection) maintains that it is "by definition uneven […] it directly depends upon the level of risk at stake for each specific processing operation."[124]

Enforcing rights has normally been a reactive process – the problem arises before being subject to sanctions. The risk-based approach is by definition proactive.[125] A central part of proactive law is for subjects to know their obligations, so as to avoid legal issues.[126] It consists, firstly, of a promotive dimension – the encouragement of lawful behaviour. Secondly, a preventive dimension – the inclusion of sanctions and other deterrents. Essentially, "[t]he *Proactive Law* approach is focused on *success* rather than failure."[127] Therefore, as Mahler maintains, "the *proactive* function of the law is essential for the understanding of the relationship of law and risk management."[128]

The act of regulating is the implementation of compliance tools, including monitoring and enforcement, designed to steer behaviour in the direction of a set goal, i.e. law, industry standards, social norms, or the rules of badminton.[129] Regulatory law can be viewed as separate from private law. Private law is enforced ex-post, i.e. through court proceedings, whereas regulatory law aims to impose standards ex-ante, i.e. product regulation.[130]

Both the rights- and risk-based approach operate with a principle of proportionality,[131] however the two do not necessarily implement the principle in the same way.[132] For the risk-based approach, proportionality "can be conceptualised of terms of two balancing tests associated with risk mitigation measures."[133] Gellert identifies the first as a test of legitimacy – whether the adopted measures pursue a goal, which is legitimate in and of itself. Secondly, whether the adopted measures are tolerable or acceptable. In addition, Gellert points to the necessity-test. In essence, aiming to "diminish the impact of the adopted measures."[134] To put it differently: impose as little as possible, whilst achieving as much as possible.

[123] Koene (2021) p. 10.

[124] (2020) p. 3.

[125] Mahler (2010) p. 7.

[126] EESC (2008) p. 2.

[127] Ibid. p. 8.

[128] Mahler (2010) p. 66.

[129] Gellert (2020) p. 41.

[130] Ibid. p. 44.

[131] Ibid. p. 10.

[132] Ibid. p. 15.

[133] Ibid. p. 12.

[134] Ibid. p. 15.

As mentioned above, a *traditional* conception of law consists of yes/no questions of legality – also referred to as "command and control" regulation.[135] Another form of regulation is meta-regulation – principles-based.[136] Rather than a detailed, prescriptive approach to regulation, meta-regulation defines a desired outcome, without sketching out how to get there. By leaving the regulatee's to define the individual steps towards the regulatory goal, the meta-approach leads to more freedom, but also an increase in responsibility, as the regulatee has to identify the appropriate steps.[137] From the regulator's point of view, the flexibility involved means that the "management of risk can be delegated to the involved parties" whilst the regulator can still maintain "a certain degree of control over the risky activities."[138] The amount and nature of legal obligations, i.e. implementing a risk management system is often predicated on the object being regulated, i.e. a specific product "[contexts] where the authorities have identified specific risks."[139]

There are certainly issues with the risk-based approach. An obvious issue is how to measure the *value* of outcomes associated with the manifestation of a risk. If it can be measured as a monetary loss, the exercise is simple. However, if the outcome is an adverse effect to fundamental rights, it becomes rather difficult.[140] Gellert points to another aspect, the redefining of compliance.[141] The rights-based approach of compliance means following each provision to the letter, whereas the risk-based approach means the creation of instruments to fulfil a general principle. To summarize "[c]omplying through risk-based instruments is simply not identical to complying from a "legal" viewpoint."[142] The risk-based approach also entails a collaborative mode of regulation where the responsibilities of the parties involved, namely regulator and regulatee, are easily blurred.[143] Furthermore, the approach might be shrouded in an illusion of simplicity and cost-effectiveness, as the implementation of a framework "can be resource intensive, very complex, and inconsistent."[144] With these general observations in mind, the question is how a risk-based approach works in practice.

[135] Ibid.

[136] Ibid. p. 20.

[137] Ibid.

[138] Mahler (2010) p. 61.

[139] Ibid. p. 68.

[140] Ibid. p. 39.

[141] (2020) p. 23.

[142] Gellert (2020) p. 23.

[143] Ibid.

[144] Ibid.

5.3 Risk management

As we have seen, the risk-based approach to regulation is predicated on identifying an acceptable level of risk. It is then up to the regulatee to implement measures for the management of said risk: "[r]isk management can thus be seen as the tools, processes, and methodologies that implement and actualise the otherwise abstract technique of risk (ie how to concretely assess and manage risks)."[145] In a purely technical sense, risk management consists of a given set of steps, with further sub-steps. The process is divided into two main steps. Firstly, the measurement of risk – the risk assessment. Secondly, the concrete steps/measures implemented to manage said risk – risk management *sensu stricto*.[146]

Within the risk management nomenclature, the following three terms are central: (i) the risk/event – what actually occurs, (ii) risk factors – what contributes to the manifestation of the risk and their likelihood, (iii) the consequences – the harms or benefits resulting from the manifestation of a risk.[147] When it comes to the actual assessment, there are some divergences between different authors. For instance, Mahler refers to sources, rather than risk factors.[148] Furthermore, there is an obvious difference in both how many, and the contents of each step in a risk assessment. Gellert presents a model for risk assessment which consists of three steps. First, the assessor must evaluate the "risk criteria" – namely the risk factors and consequences.[149] Second, the "risk identification", which applies the risk criteria to the risk, thus determining whether it is sufficiently risky.[150] Third, if the first two steps reveal a risk, assessing/analysing the magnitude – how likely is the risk to manifest, and how severe are the consequences.[151] The assessment results in the designation of a risk category or classification. Importantly, assigning a high risk within the structure of risk management does not directly correlate to whether the AI system has a high-risk classification.[152]

There are different approaches to risk assessment, namely toxicology, and epidemiology. In risk management theory, toxicology is concerned with the level of risk (severity and probability).[153] It follows that toxicology can be understood as the assessment of the "risk" and the "risk

[145] Ibid. p. 28.

[146] Ibid.

[147] Ibid. pp. 29-30.

[148] (2010) p. 41.

[149] Gellert (2020) p. 30.

[150] Ibid.

[151] Gellert (2020) p. 30.

[152] Mahler (2022) p. 255.

[153] Gellert (2020) p. 32.

factors". Epidemiology on the other hand, is concerned with the effects "consequences" of a risk.[154] Gellert clarifies this distinction with the following example from data protection:

> *[I]f a data processing operation is considered a risk, [...] Toxicology will address the processing as risk as such (which is thus the source of harms), and will therefore include such elements as the nature and type of data, the type of processing operation, its scope, its context, the status of the controller and the data subject, etc. Epidemiology will assess the harms in themselves and the way they affect data subjects (ie the risk targets). Such harms may include discrimination, defamation, loss of bargaining power, distress, irritation, fear, etc.*[155]

Risk management *sensu stricto* consists of two steps. First, "a decision as to whether or not to take a risk."[156] Second, "measures to reduce the risk once it has been decided to take it [risk mitigation]."[157] Risk management *sensu stricto* is concerned with the mitigation of risk to an acceptable level, not to eliminate it – to set a standard.[158] The types of measures implemented in risk mitigation can be divided into active/corrective measures, and passive/preventative measures. Corrective measures attempt to lower the probability and magnitude. Preventative measures are concerned with the effects/consequences of the risk.[159] This distinction becomes clear if considered in terms of temporality. The corrective measures are risk management ex-ante, the preventative are ex-post.

5.4 Summary

Risk, risk-based regulation, and risk management theory are related concepts. Yet, words might have different meanings depending on the topic. I therefore find it necessary to provide a brief summary.

Firstly, a risk is an event measured in risk factors, with positive or negative consequences. The risk-based approach to regulation imposes legal obligations depending on the level of risk. The risk-based approach might involve classification depending on the level of risk. Risk management consists of an assessment of risks, and the implementation of measures to limit the consequences, by reducing the size of risk factors.

[154] Ibid.

[155] Ibid. p. 33.

[156] Gellert (2020) p. 29.

[157] Ibid.

[158] Ibid. pp. 33-34.

[159] Ibid. pp. 34-35.

As we have seen, assessing a risk is a systematic process. The following figure is a simplified way of representing the assessment of risk. Each risk is assessed in terms of likelihood, and expected impact. Likelihood x impact equals a risk category.

Very likely	🟨	🟧	🟥	🟥
Likely	🟨	🟨	🟧	🟥
Possible	🟩	🟨	🟨	🟧
Unlikely	🟩	🟩	🟨	🟨
	Minimal Impact	Moderate impact	High impact	Very high impact

Fig. 3 Basic Risk Matrix

6 Risk management in the Artificial Intelligence Act

6.1 Introduction

So far, this thesis has explored the contents of central provisions in the Artificial Intelligence Act, risk-classification, risks associated with AI, and risk management theory. The question remains how this all works in practice. It is clear that high-risk systems "should only be placed on the Union market or put into service if they comply with certain mandatory requirements."[160] This section presents the requirements for risk assessment, and measures for risk management included in the AIA. As mentioned earlier in this thesis, many of the risks associated with AI are still unexperienced. Yet, as maintained in the preamble, "requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights."[161] Except for requiring a risk management system, the AIA only loosely tackles the question of assessment and management.

Article 8(1), AIA stipulates that high-risk AI systems shall comply with the requirements in Chapter 2 – Requirements for high-risk systems. In order to ensure compliance, the intended purpose, and the risk management system, referred to in Article 9, shall be taken into account. Chapter 2 contains requirements for good data governance, technical documentation, record-keeping, transparency and provision of information to users, human oversight, accuracy, robustness and cybersecurity. These are all relevant to the key provision, Article 9 – risk management system.

[160] AIA recital 27.

[161] AIA recital 43.

It follows from Article 16(1)(a), that the provider has the obligation to ensure compliance with the requirements, including implementing such a risk management system. Both the EESC and EDPB-EDPS are of the opinion that all high-risk systems should undergo a third-party risk assessment, ex-ante.[162] Furthermore the EESC questions whether the measures included in the AIA are at all sufficient to mitigate the risks associated with AI.[163] They point out that the 5 requirements in the AIA may not be sufficient to mitigate risks associated with "less mentioned fundamental right […such as] human dignity, the presumption of innocence […etc]."[164]

6.2 Risk management system

Article 9(1) requires that a risk management system shall be established, implemented, documented and maintained for all high-risk systems. Secondly, Article 9(2) requires that "[t]he risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic updating." The management system shall consist of the steps listed in Article 9(2)(a)-(d).

Article 9(2)(a) requires identification and analysis of known and foreseeable risks associated with each high-risk AI system. This implies two distinct processes – 'identification' and 'analysis'. The preamble does not mention how to interpret either. In its usual meaning, to identify is to recognize or distinguish. Taken at face value, the provision therefore indicates that the provider should actively find out which risks are associated with the system. An analysis means a methodical examination. Seen in conjunction, 'identification' and 'analysis' translates to the 'assessment' of risks.

Article 9(2)(a) distinguishes between the 'known' and the 'foreseeable'. This indicates that the provider of the system has to both accrue knowledge about similar systems, and as part of the assessment of the system, imagine which risks might become relevant in the future.

Seen through the lens of risk management, Article 9(2)(a) thereby stipulates that the provider has to (i) identify risks, (ii) conduct a risk assessment.

Article 9(2)(b) requires the 'estimation' and 'evaluation' of the risks that 'may emerge' when the high-risk AI system is used in accordance with its 'intended purpose' and under conditions of 'reasonably foreseeable' misuse. The provision is confusingly similar to Article 9(2)(a). 'Estimation' and 'evaluation' would indicate the exact same process as 'identification' and 'analysis'. However, it refers to risks that 'may emerge' when the system is used. This indicates that Article 9(2)(a) requires risk assessment ex-ante, whereas Article 9(2)(b) requires risk assessment ex-post.

[162] EDPB-EDPS (2021) p. 32, EESC (2021) p. 2.

[163] (2021) p. 3.

[164] Ibid. p. 6.

In their draft opinion, JURI opted to remove Article 9(2)(b), and amend 9(2)(a) to a more intelligible version: “identification and analysis of the known and foreseeable risks of harms most likely to occur to the health, safety or to the fundamental rights in view of the intended purpose of the high-risk AI system.”[165]

Article 9(2)(c) requires further ‘evaluation’ of other risks revealed by the post market monitoring system in Article 61.[166] Essentially re-iterating the requirement for a ‘continuous iterative process’.

In summary, Article 9(2)(a)-(c) requires that providers of AI systems should assess the risks associated with their system from early development, before it is placed on the market, and whilst in use.

According to Article 9(2)(d) providers of high-risk AI systems should adopt suitable risk management measures in accordance with Article 9(3)-(9). Whereas Article 9(2)(a)-(c) required the assessment of risks, this is a requirement for risk management *sensu stricto*. The provisions in Article 9(3)-(9), AIA are a combination of actual, and requirements for, risk management measures. I will limit the discussion to Article 9(4).

Article 9(4) states that the risks management measures should leave ‘residual’ risks associated with each ‘hazard’, and the overall ‘residual’ risks associated with the high-risk AI systems at a level that is ‘judged acceptable’ if the system is used within its intended purpose or foreseeable misuse. Firstly, ‘residual’, means that the measures do not have to eliminate a risk in its entirety. The use of the words ‘risk’ and ‘hazard’ are ambiguous.[167] The wording of the provision “risks associated with each hazard” could be interpreted as “the likelihood of each risk”, it could also mean “the likelihood of negative consequences”, or “the risks associated with each risk”. The “overall risk of the systems” leads to more doubt still. Firstly, it is unclear what an ‘overall risk’ is. The plural: systems, leads to the question of whether the provision indicates that the risk management system should mitigate risks associated with all systems, or one system in particular. Lastly, the provision does not specify who should ‘judge’ what is ‘acceptable’. The seemingly best way to interpret the provision is that the risk management measures should mitigate the risks associated with the high-risk AI system, to a level which insures compliance with the regulation.

Article 9(4)(a)-(c) provides further criteria for identifying the most appropriate risk management measures. Firstly, elimination or reduction of risks through ‘adequate’ design and development, cf. Article 9(4)(a). The provision duly reflects the requirement for the risk management system to be in place throughout the entire lifecycle of the system. ‘Adequate’ implies that the

[165] (2021) p. 47.

[166] Article 61 will not be discussed in this thesis.

[167] See section 5.1, with reference to Mahler (2022) pp. 259-263.

requirements for design and development depend on the level of risk. The suggested amendment by JURI specifies that risk management through design and development should be limited to what is “commercially reasonable and technologically feasible”.[168] Article 9(4)(b) further requires adequate mitigation and control measures where risks cannot be eliminated. Finally, adequate information to the user, pursuant to Article 13, cf. Article 9(4)(c).

Finally, as stipulated by Article 9(4) eliminating or reducing risks, ‘due consideration’ shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used.

As a framework for risk assessment and risk management, Article 9, AIA can thereby be broken down as follows.

Firstly, find known and foreseeable risks associated with the high-risk AI system when used, either as intended, or under foreseeable misuse, within the intended area, cf. Article 9(2)(a).

Secondly, the risk assessment:

Evaluate the risk criteria – the risk factors (what contributes to the manifestation of the risk, and their likelihood) and their consequences (positive or negative), cf. Article 9(2)(a)-(c).

Risk identification – apply the risk criteria to the risk, to determine if it is indeed sufficiently risky, cf. Article 9(2)(a)-(c). I.e. whether mitigated to an acceptable level, cf. Article 9(4).

Assessing magnitude – likelihood, and severity of consequences, cf. Article 9(2)(a)-(c).

Thirdly, implementing risk management measures – risk management *sunsu stricto*:

Determine whether the risk is worth taking. If yes, implement mitigating measures.

Active/corrective measures – lowering likelihood and magnitude, cf. Article 9(4)(a) and (c).
Passive/preventive measures – alleviating the consequences, Article 9(4)(b)-(c).

The remainder of Chapter 2 gives further indication of what a risk assessment should include, and possible mitigating measures, in order to comply with the AIA. These are the requirements for high-risk systems listed in Articles 10 through 15. The following discussion is limited to a select few of these provisions.

6.3 Data and data governance

Article 10, AIA lists criteria for good data, and data governance. It follows from Article 10(1) that the provision applies to systems that use techniques involving the ‘training’ of models with ‘data’. Article 3(29) defines ‘training data’ as “data used for training an AI system through fitting its learnable parameters.” This refers to the input. As illustrated in section 3, the training

[168] JURI (2021) p. 50.

data for a supervised learning system consists of labelled inputs. In order to ensure good data governance, Article 10(1) further stipulates that datasets should meet the quality criteria in paragraphs 2 through 5. The quality of the data is, as stated in the preamble "essential for the performance of many AI systems […] to ensure that the high-risk AI system performs as intended and safely."[169]

Article 10(2)(a)-(g) lays out data governance and management practices of particular concern. These include design choices, cf. (a), data collection, cf. (b), preparation, cf. (c), relevant assumptions, cf. (d), suitability assessment, cf. (e), control for bias, cf. (f) and, find gaps and shortcomings, cf. (g). These practices of particular concern are indicative of how data governance becomes part of the overall risk management – they stipulate a form of best practice for risk assessment, and indicate possible mitigating measures.

As an example, Article 10(2)(f) requires examination in view of possible biases. This means that biases are something to avoid, and that the provider should work proactively to ensure that it does not occur. If seen as a provision for risk management, Article 10(2)(f) requires that the provider must assess whether the data is a contributing factor to the risk of biases, and if so mitigate this risk by adjusting datasets accordingly.

According to Article 10(6), high-risk systems that do not explicitly involve the training of models, still has to comply with the provisions in Article 10(2).

6.4 Transparency and human oversight

Article 13, AIA requires transparency, and information to users. The worry is that "certain AI systems [are] incomprehensible to or too complex for natural persons."[170] Ensuring a minimum level of transparency and explainability is, as mentioned earlier in the thesis, a central issue with AI.

According to Article 13(1), systems shall be 'designed' and 'developed' to ensure that their 'operation' is 'sufficiently transparent' to enable 'users' to 'interpret' the system's 'output' and use it appropriately. This means that transparency should be a design feature – the systems should be inherently transparent. The 'operation' refers to the systems 'computation' – what the system actually does. Requirements for transparency and explainability are modified by the technical knowledge of the user – the system is sufficiently transparent when the user is able to interpret, and use the output. JURI's proposed amendment clarifies the meaning of interpretation as understanding "the rationale of decisions" by "generally knowing how the AI system

[169] AIA recital 44.

[170] AIA recital 47.

works, and what data it ingests."[171] In practice, this imples that users should be quite technologically adept.

Article 13 (1) further identifies that an appropriate 'type' and 'degree' of transparency shall be ensured, 'with a view' to achieving compliance with the obligations of 'users' and 'providers' in Chapter 3. 'Type' of transparency could indicate the concrete way in which the system is transparent, i.e. accompanying instructions, or a system designed to explain its operations – explainable AI (XAI). It is up to the provider to decide on 'type' and 'degree'. They are determined 'with a view', that is the goal of, Compliance with Chapter 3.

Article 13(2)-(3) specifies the primary risk mitigating measure to ensure compliance with transparency obligations – appropriate instructions. According to Article 13(2), instructions and accompanying information should be concise, complete, correct, clear, and accessible and comprehensible to the user. This means that the requirement for instructions depends on both the complexity of the system, and user competence.

Article 13(3)(b) further clarifies that the information should specify the systems characteristics, capabilities, and limitations. Among these, Article 13(3)(b)(iii) requires information on whether use, within the intended purpose or under foreseeable misuse, may lead to risks to health and safety or fundamental rights.

This means that, as part of their overall risk assessment, the provider must determine which level of transparency is most suited, taking into account the system in itself, the user of the system, and how the system is to be used. Ensuring appropriate transparency thus means the managing of a legal risk – ensuring compliance with Chapter 3. Among risk mitigating measures, the provider should include instructions.

Article 14 lays out the requirements for human oversight. Systems should be designed and developed in such a way that they can be effectively overseen by natural persons, with the aim of preventing or minimising risks to health, safety and fundamental rights, cf. Article 14(1)-(2). Article 14(4)(b) stipulates that the user should remain aware of automation bias, in particular when the system provides information, or gives recommendations. The provision identifies a risk, automation bias, and an area where it can occur, information or recommendation. Furthermore, it loosely suggests a mitigating measure, 'remain aware'.

[171] (2021) p. 62.

7 Case study for AI risk management

7.1 Introduction

In February 2020, The Hague District Court delivered judgement in the case of Nederlands Juristen Comité Voor De Mensenrechten (NJCM) v. The Netherlands, regarding the Systeem Risico Indicatie (SyRI).[172] By utilising the SyRI case as an example, this section illustrates the relationship between the risk-based approach to regulation in the Artificial Intelligence Act and risk management. The purpose of the case study is to take on the role of the provider, and demonstrate elements of a risk management system, pursuant to Article 9, AIA. With reference to the provisions in Chapter 2 – Requirements for high-risk systems, I attempt to assess, and suggest mitigating measures for the risks associated with the use of SyRI, as they are presented in the judgement.

The SyRI case involved a system used to determine access to public services, including revoking such services. It made individual risk assessments, based on a risk model, in order to determine the likelihood of a natural person committing, or having committed a criminal offence. For the purposes of the case study, I assume that the system would classify as a high-risk AI system, in accordance with Article 6(3), AIA, cf. Annex III, point 5 (a), and point 6 (a). Regarding systems determining access to public services, the AIA notes that the "Regulation should not hamper the development and use of innovative approaches in the public administration […] provided that those system do not entail a high risk to legal and natural persons."[173] The question was whether the system did indeed entail such a risk.

Without access to the underlying algorithms, training models, or statistics there is no way to verify the findings – they are general assumptions. It is vital to note that the analysis is for illustrative purposes only, to illustrate *how* such an analysis might be undertaken, not to draw factual conclusions about the SyRI system. Suggested risk mitigating measures are discussed in light of the assumptions made in the risk analysis, in order to meet the requirements in Chapter 2, AIA.

7.2 The SyRI judgement

SyRI was a system utilised by the Dutch government to detect fraud, including tax, and social benefits. The system was characterised as a "technical infrastructure with associated procedures with which data can be linked and analysed anonymously in a secure environment, so that risk reports can be generated." The purpose of which was to determine whether to conduct further investigation.[174] Data from governmental and other bodies were shared, and fed to the model.

[172] Case number C-09-550982-HA ZA 18-299 (English), hereafter SyRI.

[173] AIA recital 37.

[174] SyRI pt. 3.1-3.2.

Between 2008 and 2014, SyRI and its precursor was utilised in a so-called neighbourhood-approach, meaning that the system was deployed in a select few geographical areas.[175]

The main question before the Court was whether the legislation regulating SyRI was in violation of Article 8, of the ECHR, Articles 7 and 8, of the Charter of Fundamental Rights, and several provisions in the GDPR. The court questioned whether the legislation struck a fair balance between the benefits of new technologies, and the respect for private life. The Court did find a breach of Article 8 ECHR.

A second question, was whether the government had to disclose the risk models, and indicators used in related projects, G.A.LO.P. II and Capelle.[176] The risk models and indicators included the underlying algorithms, instructions etc. Insight into these models would have helped in assessing whether the system made, or was trained to make biased decisions. The Court dismissed this claim.

7.3 Risks associated with the system

The first step in the risk management system is to identify and analyse known and foreseeable risks, cf. Article 9(2)(a), AIA. As mentioned earlier, to 'identify' risks involves finding out which risks are associated with the system, and systems like it. The risks that need to be identified, are those associated with the AI system itself – risks that stem directly from the technology or its application.[177]

It was clear that SyRI involved the collection of large amounts of data, that the datasets underwent pseudonymisation, and that subsequent identification required the correct decryption key.[178] The Court assessed whether SyRI was an example of 'big data', or utilised 'deep learning'. The plaintiff argued that SyRI was an example of "unstructured and random automated linking of files of large groups."[179] The Advisory Division of the Council of State noted that the program "run different types of files […] against each other […] in line with the use of deep learning and self-learning systems […] investigating as many links as possible without preconceived notions."[180] Regarding self-learning systems, they note that "[t]hey can therefore not substantiate their predictions in a legally sound manner," and that "[a]n administrative organ that partially bases its actions on such a system is unable to properly justify its actions and to

[175] SyRI pt. 3.9.

[176] Collectively referred to as SyRI.

[177] Precluding i.e. the question of whether there was a breach of Article 8 ECHR.

[178] SyRI pt. 4.29-4.31.

[179] SyRI pt. 6.45.

[180] SyRI pt. 6.46.

properly substantiate its decisions."[181] The Government argued that SyRI consisted of a simple decision tree, comparing previously gathered data for discrepancies. That it did not predict future behaviour, but rather built a risk model, based on those discrepancies.[182]

Based on the description in the judgement, SyRI is likely a system that utilises machine learning. As mentioned in section 3, machine learning is an automated statistical analysis of correlation and probability, leading to conclusions about the world. The plaintiff's description "unstructured and random" implies that it is developed through unsupervised learning. Such systems find correlation, but not always the right causation, meaning that the output might be inaccurate. However, according to the judgement, any use of a risk report generated by the risk model, required that the user give feedback on the results, with the intention of increasing the effectiveness of the model.[183] This implies a system developed using reinforcement learning. In addition, based on the feedback, it was possible to adjust the model. Earlier uses of systems designed to give recommendations on decisions that have legal impact have, as mentioned in section 4[184] of this thesis, resulted in discriminatory outcomes. The system is also subject to the requirements for good data and data governance, cf. Article 10(1), AIA, as it develops models based on training data.

The nature and application of the system makes algorithmic discrimination a foreseeable risk. Therefore, the risk that needs to be addressed is the risk of an illegal decision on grounds of discrimination, in violation of fundamental rights. Whether such a decision is made depends upon whether the SyRI system has flagged an individual as high risk, and secondly whether the government decides to act on that recommendation/flagging, to conduct a formal investigation.

7.4 The risk assessment

A full formal risk assessment falls beyond the scope of this thesis. I will therefore present a simplified, two-step version of a risk assessment. This will include discussing risk factors and consequences, and assessing the magnitude of risk. In cases of algorithmic discrimination, two key factors often lie at the core, the datasets, and automation bias.

The plaintiff argued that, "given the large amounts of data that qualify for processing in SyRI, including special personal data, and the circumstance that risk profiles are used, there is in fact a risk that SyRI inadvertently creates links based on bias, such as a lower socio-economic status

[181] SyRI pt. 6.46.

[182] SyRI pt. 6.47-6.48.

[183] SyRI pt. 4.32.

[184] See Angwin et.al. (2016), Hannah-Moffat (2019), Russell (2022).

or an immigration background."[185] Article 5a. 1 paragraph 3 SUWI Decree of the SyRI regulation listed the numerous data-categories available for processing.[186] These data have to be examined in view of possible biases, cf. Article 10(2)(f), AIA.

For this analysis, I will only consider six of the seventeen categories allowed for processing under the SyRI regulation. These are work data, tax data, property data, identifying data (i.e. address, age, gender), education, social benefits. Data is not biased in and of itself, the question of whether it contributes to the risk, depends on how the system weighs different data points. If the SyRI system was fully unsupervised, then it is impossible to know how much weight it attributed to each data point. If however, the system was informed whenever a decision to pursue investigation was made, then it would involve reinforcement. Without access to the underlying algorithm, it is necessary to make some general assumptions. Based on the information that each use of a risk report required feedback, it is likely that reinforcement learning was part of the system.

As the purpose of the system was to construct a risk model to determine the likelihood of fraud, it is plausible, or at least ideal, that data points were given different weight. Data relating to taxes, and whether the person is, or has formerly received social benefits are likely the best indicators of whether fraud has been committed. In able for the system to construct a risk model with any accuracy, these data would surely have to be included. The system might find correlation between prevalence of fraud and income groups. However, the likelihood of this leading to discrimination down the line does not seem high. As they help describe a financial situation, work, education and property data may be indicative of tax- or social benefits fraud, though not as much as the actual tax returns. Identifying data, such as address, age, or gender are likely not indicative of whether the person has acted fraudulently. In addition, if the system finds a disproportionate amount of risk of fraud in one geographical area, it might conclude that people living in that area are more likely to commit fraud. Similarly, it might conclude that fraud is restricted to certain age groups.

When considered as risk factors, data related to taxes and social benefits are unlikely to contribute to a biased recommendation. Work, education, and property data are somewhat likely to contribute to a biased recommendation. Identifying data are likely to contribute to a biased recommendation.

A biased recommendation made by the system does not directly lead to an illegal decision. However, if the system receives some sort of reinforcement learning, accepting the recommendation would lead to more biased recommendations in the future. The consequence of a biased

[185] SyRI pt. 6.92-6.93.

[186] SyRI pt. 4.17.

recommendation is therefore a system trained to give more biased recommendations. As mentioned, authorities had to give feedback each time they asked for a risk report. Thus, it is likely that, if indeed the recommendations were biased, they would become increasingly so.

This leads to the deciding factor, whether the recommendation to investigate was followed. According to the judgement, “[a] risk report means that a legal or natural person is deemed worthy of investigating”[187] The judgement does not indicate how often a risk report lead to an investigation, however it does inform that “[i]n 19 of the 21 completed intervention team projects a so-labelled neighbourhood-oriented approach had been applied.”[188] Meaning that substantial effort had been aimed at specific geographical areas, which were identified as ‘problem areas’. This could indicate that the model had begun associating fraud with not just relevant data, but also other categories.

Based on the amount of available data, the resulting investigations, and the possibility of a positive feedback loop, the judgement does indicate that the SyRI system might indeed have lead to a form of algorithmic discrimination. If the Government utilised a machine learning system that perpetuated societal biases, and lead to a discrimination, the impact would undoubtable be very high.

The risk of algorithmic discrimination can thereby be placed in a risk matrix. The category ‘orange’ can be designated as high risk.

Very likely	🟨	🟧	🟥	🟥
Likely	🟨	🟨	🟧	🟥
Possible	🟩	🟨	🟨	🟧 Discrimination
Unlikely	🟩	🟩	🟨	🟨
	Minimal Impact	Moderate impact	High impact	Very high impact

7.5 Risk management measures

Once the risk of algorithmic discrimination has been assessed, and found to have a high level of risk, the question is how to mitigate it. The risk management system should adopt suitable risk management measures, cf. Article 9(2)(d), AIA. These should eliminate, or reduce the risk to an acceptable level. Mitigating measures should be implemented through design and development, or consist of adequate mitigation and control mechanism, or both. In addition, the user should be given adequate information, cf. Article 9(4).

[187] SyRI pt. 3.2.

[188] SyRI pt. 3.9.

The first risk mitigating measure to consider is good data and data governance. The risk assessment showed that some of the data available for processing could lead to biases. Requirements for good data governance therefore imply that the provider of the system has to remove some of the data categories, i.e. identifying data. This both reduces the likelihood of algorithmic discrimination, as well as ensuring compliance with Article 10, AIA. However, there is no guarantee that all bias can be eliminated. Therefore, reviewing the datasets is a risk mitigating measure that, through design and development, reduces a risk as far as possible, cf. Article 9(4)(a).

Secondly, the system has to be transparent, and provide information to the user, cf. Article 9(4)(c) cf. Article 13. The risk report consisted of "individualised information from [SyRI] containing a finding of increased risk" where "coherently presented data from [SyRI] forms part".[189] The Government claimed that the risk model was based on a simple decision tree – a feedforward network. If this is indeed the case, interpreting the output would require little technical skill. If the risk models were complex algorithms involving machine learning, utilising large quantities of data, then meeting the requirements for transparency and explainability would be more cumbersome. As mentioned in section 6, JURI's suggested amendment required the user to generally know how the system worked. For advanced systems, meeting the requirements might therefore involve the development of separate systems for interpretation, or XAI. In light of the intended purpose, and impact of SyRI, these two options are both reasonable requirements for adequate risk mitigation.

Lastly, and most importantly, pursuant to Article 14(4)(b), the system should have built in human oversight mechanisms, that makes the person responsible for human oversight to remain aware of automation bias.

8 Final remarks

One of the main problems regulators will be faced with during development of AI regulation is the definition of AI. At this point, it is unclear whether it is possible to find a legal definition that correlates with a scientific one. Though Article 3(1), AIA, has been subject to criticism, it might be in the public interest for forthcoming regulation to encompass more than a strict scientific definition of AI – allowing the regulation of a wider set of systems. For instance, it is far from certain that the SyRI system would qualify as a high-risk AI system, even under the current definition in Article 3(1).

For AI regulation, the risk-based approach to regulation is likely preferable to a rights-based one. The option to update lists, guidelines, and accompanying annexes allows a level of flexibility, unlikely in the case of a strict blanket-regulation. The approach in the current iteration of the AIA is one of several different risk assessments. This thesis has only touched on two of

[189] SyRI pt. 4.12.

these, the abstract risk assessment in Article 7, and the concrete risk management system in Chapter 2. As the development of AI continues, lists of high-risk systems will likely see many an update. A question that remains un-answered is whether the ex-post assessment in Article 7 will be able to keep pace with rapid technological development.

This thesis provides a simplified illustration of requirements for a risk management system in the AIA, and AI risk management in practice. A central question for this thesis was whether risk management could be effectively combined with legal reasoning. The requirements for high-risk AI systems in Chapter 2 are, for the time being, an unorderly combination of technical terms from risk management, and ambiguous formulations. Opinions from parliamentary committees, particularly the Committee on Legal Affairs, point in the direction of an Artificial Intelligence Act that incorporates terms from risk management into the wording and structure of the relevant provisions, making it a true risk-based approach.

AI is in its infancy. How to define, regulate, and manage these technologies is a big unknown. One of the great questions for the future is what to do if/when a machine becomes so intelligent that it is indistinguishable from, or more likely smarter than, a human – the question of 'robot rights'. A possible approach is to ask, as Bentham did, "Can they suffer?" Another is to avoid the discussion entirely, by never creating human-like AI in the first place.[190] I will allow myself to end this thesis with these differing views on the prospect of 'the technological singularity'.

> *Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligent explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.*[191]

> *The singularity will allow us to transcend these limitations of our biological bodies and brain. We will gain power over our fates … We will be able to live as long as we want … We will fully understand human thinking and will vastly extend and expand its reach. By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence.*[192]

[190] Russell (2022) p. 1052.

[191] I. J. Good (1956) in Russell (2022) p. 1055.

[192] Ray Kurzweil (2005) in Russell (2022) p. 1056.

9 Table of reference

9.1 EU-regulation (including proposals)

The Charter	OJ C 364/1, 18/12/2000. Charter of Fundamental Rights of the European Union.
TEU/TFEU	OJ C 326/1, 26/10/2021. Consolidated versions of the Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU).
GDPR	Regulation (EU) 2016/679 of The European Parliament and of The Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).
AIA (April proposal)	COM(2021) 206. ‘Proposal for a Regulation of The European Parliament and of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.’ (21st April 2021).
AIA	COM(2021) 206. ‘Proposal for a Regulation of The European Parliament and of The Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.’ Presidency compromise text (29th November 2021).

9.2 EU opinions, guidelines, communications, and recommendations

EESC	INT/415-EESC-2008-1905. Opinion of the European Social and Economic Committee on the proactive law approach: a further step towards better regulation at EU level (own initiative opinion) (3rd December 2008)

Art 29 WP	Article 29 Data Protection Working Party. Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is “likely to result in a high risk” for the purposes of Regulation 2016/679 (4th October 2017)
AI for Europe	COM(2018) 237 final. Communication from the commission. Artificial Intelligence for Europe. (25th April 2018)
White Paper on AI	COM(2020) 65 final. White Paper. On Artificial Intelligence – A European approach to excellence and trust. European Commission (19th February 2020)
EUCO 13/20	Special meeting of the European Council (1 and 2 October 2020) – Conclusions
2020/2021(INL)	European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies. (20th October 2020)
11481/20	114841/20. Presidency conclusions – The Charter of Fundamental Rights in the context of Artificial Intelligence and Digital Change (21st October 2020)
EDPB-EDPS	Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (18th June 2021)
EESC	EESC-2021-02482. Opinion. European Economic and Social Committee. AI/Regulation. COM(2021) 206 final - 2021/0106(COD). (22nd September 2021)

JURI Draft Opinion of the Committee on Legal Affairs for the Committee on the Internal Market and Consumer Protection, and the Committee on Civil Liberties, Justice and Home Affairs on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts. COM(2021) 0206 - 2021/0106(COD) (2nd March 2022)

9.3 Case Law

SyRI. Rechtbank Den Haag. Case number: C/09/550982/HA ZA 18-388. Nederlands Juristen Comité Voor De Mensenrechten, et. al. vs. The State of the Netherlands (Official English Translation). Reference: ECLI:NL:RBDHA:2020:1878.

9.4 Literature

Gellert, Raphaël. *The Risk-Based Approach to Data Protection*. 1st ed. Oxford: Oxford University Press, 2020.

Mahler, Tobias. *Legal Risk Management. Developing and Evaluating Elements of a Method for Proactive Legal Analyses, With a Particular Focus on Contracts.* 1st ed. Oslo: University of Oslo, 2010.

Mahler, Tobias. “Between Risk Management and Proportionality: The Risk Based Approach in the EU’s Artificial Intelligence Act Proposal.” In *Nordic Yearbook of Law and Informatics 2020-2021: Law in the Era of Artificial Intelligence*. Colonna, Liane, & Stanley Greenstein ed. Visby, Sweden: Stiftelsen Juridisk Fakultetslitteratur & The Swedish Law and Informatics Institute, 2022. pp. 247-269.

Russell, Stuart J., Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th ed, Global Edition. Harlow: Pearson Education Limited, 2022.

Van Roy, V., Rossetti, F., Perset, K., Galindo-Romero, L. *AI Watch - National strategies on Artificial Intelligence: A European perspective, 2021 edition*. EUR 30745 EN, Publications Office of the European Union, Luxembourg, 2021.

9.5 Journals and articles

Benjamin, Misha, Kevin Buehler, Rachel Dooley, Peter Zipparo. “What the draft European Union AI regulations mean for business”. *McKinsey Analytics*. McKinsey & Company, 2021.

Hannah-Moffat, Kelly. “Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates”. *Theoretical Criminology*. Vol. 23, No. 4, 2019. Pp. 453-470

Koene, Dr. Ansgar (ed.). “A Survey of AI Risk Assessment Methodologies: The Global State of Play and Leading Practices Identified”. *EY Trilateral Reasearch*. 2021

Shannon, Claude E. “Programming a Computer for Playing Chess”. *Philosophical Magazine*. Ser. 7, Vol. 41, No. 314, 1950.

9.6 Online

Angwin, Julia, Jeff Larson, Surya Mattu, Lauren Kirchner. “Machine Bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *ProPublica*. May 23rd, 2016. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed 04.03.22)

Baker, Harry. “How many atoms are in the observable universe?” *Live Science*. July 10th, 2021. Available at: https://www.livescience.com/how-many-atoms-in-universe.html (Accessed 01.02.22)

“A European approach to artificial intelligence.” Available at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (Accessed 05.01.2022)

“Regulatory framework proposal on artificial intelligence.” Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (Accessed 25.01.2022)

9.7 Audio/Video

Jon Bing Memorial Seminar. *Towards a New Legal Framework for AI in Europe: Assessing the European Commission’s proposed AI Regulation*. [Video] (25.05.2021) https://www.jus.uio.no/ifp/om/organisasjon/seri/arrangementer/2021/the-jon-bing-memorial-seminar-2021.html