Risk management considerations for artificial intelligence business applications

Gergő Barta* and Gergely Görcsi

Doctoral School of Management and Business Administration,
Szent István University,
Páter Károly utca 1, 2100, Gödöllő, Hungary
Email: Barta.Gergo@phd.uni-szie.hu
Email: Gorcsi.Gergely@nisz.hu
*Corresponding author

Abstract: The number of projects and the amount of investment into artificial intelligence (AI) based business process automation is increasing that is also due to research advancements in corresponding fields. To utilise its true power, business organisations shall identify and treat risks arising from AI, that must be reduced to an acceptable level to maintain fraud-free business operation in alignment with external legislative requirements. If risks are not assessed, then AI might cause greater headache resulting in expensive implementation without business benefit. The objective of the paper is to analyse the nature of risk elements that AI can bring to the life of corporations and the countermeasures that shall be implemented by analysing general IT risk assessment processes and the stages of intelligent system development. The article also examines frameworks for AI risk management approaching risks associated with intelligent decision making by providing guidelines of required business processes to be implemented.

Keywords: artificial intelligence; machine learning; IT risk assessment; risk management framework; business process automation.

Reference to this paper should be made as follows: Barta, G. and Görcsi, G. (2021) 'Risk management considerations for artificial intelligence business applications', *Int. J. Economics and Business Research*, Vol. 21, No. 1, pp.87–106.

Biographical notes: Gergő Barta is a Researcher and PhD student at Szent István University, Hungary in the topic of machine learning model development focusing on ensemble techniques. During his career, he was an advisor of several international companies on developing IT risk assessment frameworks.

Gergely Görcsi is a Researcher and PhD student at Szent István University, Hungary focusing on the implementation and development of effective Business Intelligence (BI) systems and processes. During his career, he established a BI Department and implemented BI solutions for a Hungarian public company in the ICT industry.

This paper is a revised and expanded version of a paper entitled 'Assessing and managing business risks for artificial intelligence based business process automation' presented at *International Scientific Conference "Contemporary Issues in Business, Management and Economics Engineering, CIBMEE-2019"*, Faculty of Business Management at Vilnius Gediminas Technical University, Vilnius, 9–10 May, 2019.

1 Introduction

AI based business solutions gained significant popularity in recent years that can be due to several advancements in its supporting technological environment (such as data management, computer performance etc.). The exact definition of AI brings several challenges, since there are different definitions available in academy and, as well as, in industry, therefore expressing what AI exactly covers is a hard task and appears to be subjective in most of the cases. That is supported by a research work performed by MMC venture capital firm, that concluded that 40% of startup companies that are claimed to use AI in their products, do not, in fact, use AI (MMC Ventures, 2019). According to a classic definition of Russel and Norvig (2005) AI is "the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment." However, this definition today more resembles of the definition of 'reinforcement learning' that is a subcategory of "Machine Learning". A more general definition was given by Borgulya (1998) that highlights that AI is thinking and acting as a human by "modelling human problem solving". Jonas Schuett (2019) was searching for an appropriate definition from a legal perspective focusing on the goal how AI could be defined by policy makers that makes it possible to establish a legal framework and regulate the use of intelligent systems. He concluded that the term "Artificial Intelligence" shall be avoided and instead, policy makers have to adopt a risk-based approach:

- decision shall be made on the risk that has to be addressed when it comes to intelligent decision making
- analysis shall be in place to adequately address the system and its components that is responsible for particular risks emerging from AI
- define the properties of the system.

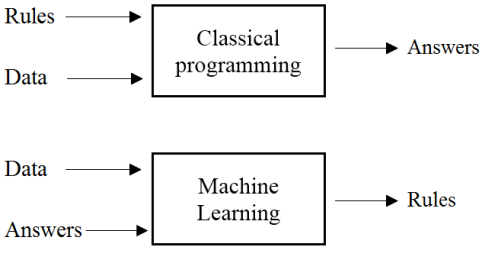
He concluded that policy makers shall not explicitly define AI as it would be misunderstood by companies utilising the technologies, the "starting point should be the underlying risk" that the set of intelligent business automation arises (Schuett, 2019). The point why it is interesting to be considered is that the use of AI can pose legislative and regulatory risks for companies, therefore the boundaries of such systems must be defined in some commonly accepted level.

The concept of AI itself exists for several decades. It was first introduced by John McCarthy that had organised a two-month work program in Dartmouth for the researchers of computer intelligence in the summer of 1956 (McCarthy et al., 1955). Back then, AI meant a computer system that was able to produce human intelligence, but AI was, in reality, a set of conditions that had led to a conclusion by evaluating logical expressions.

That was the case still in the 80's and 90's when research works were heavily focusing on building expert systems that were to solve domain specific problems.

In the late 90's Machine Learning (ML) has got more and more popularity that is the science of building intelligent systems that are capable of making decisions without explicit programming (Dua and Du, 2011). This means that no explicitly created conditions are needed in order to deduce a conclusion from an available dataset, i.e., algorithms are able to learn interrelations among data and produce the so-called intelligent decision (generate the rules automatically) that is often referred as data-driven decision making. Figure 1 represents the difference graphically between the classical programming and the machine learning paradigm.

Figure 1 Difference between the classical programming and the ML paradigm



Source: Chollet (2018)

ML is a subcategory of AI that is called AI today by industry most of the times, thus we can observe that the academic definition diverges. ML is generally divided into three subcategories by most of the scientific literature.

- Supervised learning, that has the main goal of predicting target variables that are unknown, based on the interrelations discovered from previously labelled data and explanatory variables. The methods used for supervised learning are regularly called classifiers, or in case the target variable is continuous, they can be used to solve regression tasks. These procedures are utilised in many different kinds of problems such as recognising objects from pictures or on videos, recognising speech or voices, or simply speaking, classifying different categories, or making a prediction on continuous variables e.g., prices of real estates.
- Unsupervised learning, that aims to find interrelations within the available data, however with no labelled data that the models can learn from i.e., the data structure is previously unknown. Clustering and dimensionality reduction techniques consist a part of this subcategory. Unsupervised learning methods generally appear to be useful to find patterns by grouping data with the utilisation of different distance metrics. Performing market segmentation is a popular business case.
- Reinforcement learning, that is to improve its performance by either minimising a cost function or maximising a profit function. The system is collaborating with the environment so as to achieve this goal. A classic example is a chess program or Google's Alphago that has defeated the Go game champion a few years ago (Russel, 2017).

Hereinafter, the article deems data-driven decision making (as explained above) under ‘intelligence’. The reason is that this approach eliminates the traditional software engineering concepts such as deducing a conclusion by evaluating only logical expressions i.e., the classic ‘if-else’ conditions and brings the science of creating knowledge automatically by extracting information from raw data to the forefront. The authors believe that this is the direction to create real Artificial Intelligence in the future.

The research paper, in the following sections, will analyse general IT risk assessment methodologies and general AI development processes in order to identify relevant risks for AI development and implementation by reviewing risk assessment standards and recent publications that have enlightened some of the weak points of intelligent system development. The methods and processes are evaluated step by step so that risks could be determined in the whole lifecycle of AI applications. The paper also analyses three different kind of AI risk management framework that were developed to guide companies towards a less risky adaptation of AI systems by implementing controls and business processes to mitigate arising risks.

2 Related work

To the best of the knowledge of the authors, there are not many published research work and related articles in this field as of today. Experiences show that research papers either more focus on model development to achieve high accuracy to solve different problems or on the concept of utilising intelligent applications in different business domains e.g., robotic process automation armed with AI capabilities. Business organisations are in search to find justifiable business cases for AI implementation and preparing proof of concepts rather, as it was revealed by Andrews et al. (2017), than getting ready to perform risk assessments to address risks associated with their businesses that arising from the use of AI based solutions. Nevertheless, there are some researchers who draw into attention that AI is not only the new oil, but it can endanger business reputation if projects are not managed properly. Andrew Clark examined AI development procedures in terms of audit considerations and published several articles how internal or external auditors shall adjust their audit plans when it comes to AI (Clark, 2016, 2017, 2018). Auditing is an important part of risk assessment, as the primary function of the audit team is to obtain reasonable assurance whether internal controls are operating effectively that are an option to reduce business risk to an acceptable level (Barta, 2018b). Thus, developing audit procedures to obtain assurance over internal controls is indispensable part of the organisational risk management framework and so as to appropriately assess risk, auditing must be in place and be performed in a timely manner. The authors also analysed the development processes of intelligent systems in respect of auditability, and suggested several questions that shall be addressed by auditors when testing AI applications (Barta and Görcsi, 2018). Such questions detail considerations regarding the source of data to be used, model development and evaluation criteria. Yampolskiy and Spellchecker (2016) performed a research work regarding AI failures (that are to be addressed by a risk program) and concluded that “human values are inconsistent and dynamic and so cannot be understood and subsequently programmed into a machine. Suggestions for overcoming this obstacle require changing humanity into something it is not”. This implies that AI, most probably, will never achieve its goals and eventually every AI system will fail, therefore the implementation of such systems may not be beneficial on a long run. In terms of this article, this is a relevant

consideration, as one of the risk reducing methods is to avoid risk by terminating the business process, meaning that not implementing AI will not give the opportunity to possess any risk arising from it. That means excessive risk treatment and in case of such a decision, companies have to give up any benefits AI may be able to bring.

3 Information technology risk assessment

Business organisations can hardly be imagined with no operated IT systems (applications, databases, operation systems, networks etc.) as of today, therefore IT operation is an essential business process for each company that has also brought several vulnerabilities into the life of organisations, beside its benefits. IT systems are exposed to intentional and unintentional threats and vulnerable against malicious intent if not operated in controlled environments and monitored continuously. In order to discover the nature and extent of required protection, a risk assessment shall be performed that aims to identify organisational IT assets, their vulnerabilities, relevant threats and corrective actions that must be taken so as to maintain IT assets confidentiality, processing integrity, availability, privacy, security, accountability etc. Therefore, a risk assessment is the process of uncovering existing risks and finding a solution how to mitigate them to a level that management feels comfortable with (with enforcing compliance with local, global and internal regulations, respectively).

IT risk can be deemed to be one of the components of business organisations' overall risk universe (ISACA, 2009) that includes operational risks, market risks, environmental risks, credit risks etc. There are several standards describing IT risk assessment procedures such as the ISO/IEC 27005 (2011), PCI DSS Risk Assessment Guidelines (PCI Security Standards Council, 2012), Guide for Conducting Risk Assessment (NIST, 2012), IRAM2 (Information Security Forum, 2014) etc., however, the principal approach how to conduct these assessments appears to be common in all:

- *Identify business processes*: IT risk assessments shall be closely performed in alignment with business objectives, thus identifying the relevant business processes and procedures is the first step. Business processes must be prioritised based on their criticality to the company and priorities shall be adhered to business areas that must be taking precedence. Business processes shall be detailed so that supporting IT systems can be covered in the risk assessment activity, such as the procedure of credit scoring or procurement. The starting point should, thus, always be the business.
- *Identify supporting IT assets*: IT assets shall be aligned with business procedures that means the applications, underlying databases and operation systems, network, hardware, people, locations etc., i.e., each IT asset shall be connected to a business process that they support. Manual processes shall also be considered, since hardcopy documents (such as contracts) can contain confidential information, too. Hardcopy documents are also information assets.
- *Determine business impact*: Management shall determine the business impacts of IT assets such as specifying the financial, operational, human etc. loss in case an IT asset is compromised. That generally implies that IT assets shall be categorised on a scale regarding data confidentiality, integrity and availability. The source of this

information is usually the business owners or data owners; however, they might be tempted assigning greater business impact to their own processes, therefore, an objective method must be in place.

- *Maintain a threat catalogue*: IT assets are exposed to various threats, therefore a complete and accurate list shall be created, maintained and reviewed in a timely manner, that contains relevant threats to the organisation and also they shall be aligned to IT assets to assess the impacts if an IT asset is vulnerable against a specific threat. e.g., Application are vulnerable against unauthorised access, hence 'user access review' needs to be included in this catalogue. In addition, likelihood shall be assigned to each threat that shows the possible occurrence of the particular threat. The list of threats shall be updated in regular intervals indicating the changing environment e.g., legislative. External requirements can be deemed as a business threat, too.
- *Identify vulnerabilities*: IT assets are vulnerable; therefore, their vulnerabilities shall be assessed. e.g., a system protected by a password is vulnerable against brute force attacks.
- *Calculate inherent risk*: Inherent risk is the risk that is calculated from the business criticality of an IT asset, the likelihood of relevant threat occurrence, and the vulnerability of the IT asset. Generally expressed in a qualitative manner.
- *Identify controls*: Controls are processes, rules, countermeasures etc. that has the objective to ensure that business is operating in alignment with management intention and external regulations. Controls are risk mitigating procedures and shall be identified in order to obtain a clear picture of actual risk levels, however, controls are not always operating effectively, therefore, audit procedures should be in place to assess effectiveness and for more mature companies, control efficiency.
- *Calculate residual risk*: Residual risk is calculated from the inherent risk and relevant controls. This is the risk that remains after control implementation.
- *Perform corrective action*: In case there are no controls implemented, controls are not adequate or not sufficient, then corrective actions are needed to mitigate risks. Risk treatments can be compensating controls, transferring a risk to a third party (e.g., buying insurance for critical infrastructure), accepting risk and undertaking the negative impacts if a threat exploits an IT asset vulnerability, or avoiding the risk by totally eliminating the underlying business process (no risky process, no risk) (ISACA, 2009).

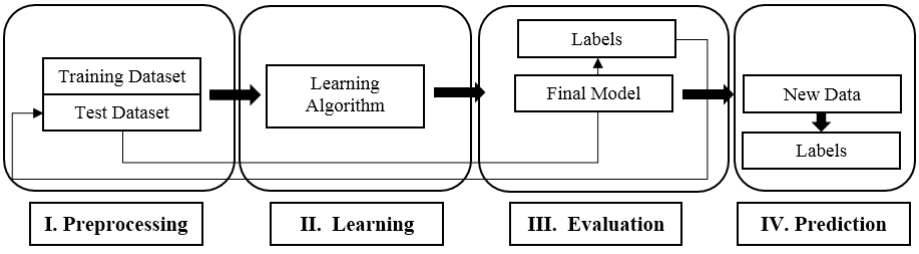
If we consider this approach as a baseline to assess risk for the IT assets, then we can conclude that intelligent IT systems shall be assessed similarly as traditional IT assets. That sounds logical. The difference in the details are the threats, vulnerabilities and controls that must be extended in order to perform a risk assessment on intelligent systems. New methods are emerged in IT capabilities, therefore different threats and vulnerabilities appear, new threats and vulnerabilities can be compensated by new controls. The business objectives, business impacts and calculating the risk ratings should remain the same, the change only occurs in the development and used methods in the IT assets. With that

in mind, assessing the risk for intelligent systems, firstly the threats and vulnerabilities must be determined, and then the mitigating controls that reduce the risks (that is the topic of the next section).

4 Machine learning general application development process

With the main goal to reveal relevant threats for intelligent applications, a general system development and operational process shall be reviewed and analysed where the processes are different from traditional software engineering. Based on the work of Raschka (2015), the development of intelligent applications is represented on Figure 2.

Figure 2 ML development phases



Source: Raschka (2015)

Raschka (2015) determines four different system development phases.

- *Preprocessing*: Data manipulation including the data transformation (feature extraction and scaling) that can differ in case of the selected algorithms, feature selection meaning the process of selecting features that are contributing the most to the final decision or prediction, dimensionality reduction that is a technique to reduce the number of features by compressing them into fewer variable e.g., principal component analysis, and the sampling that includes the separation of the dataset into different subsets such as training and test set.
- *Learning*: The selected algorithm is processing the training set and learns the interrelation among the data (the learning representation is different among algorithms). This also includes the appropriate model selection for the problem to be solved.
- *Evaluation*: The model is tested and analysed whether it performs according to the requirements e.g., capable of predicting the target variable. There should be a feedback loop in this phase, i.e., if the system does not perform as intended (the accuracy is not in alignment with a predefined value), then the whole process shall be repeated that might result in new data gathering, new data transformation, new algorithm selection etc.

- *Prediction*: The system is implemented into production and used for its intelligent decision-making purposes, hopefully, as management intended. Since these algorithms are not only capable of decision making, but they are able to perform other cognitive tasks, too, e.g., they can produce new content (e.g., images, text, speech by the use of intelligent algorithms such as Generative Adversarial Networks and Autoencoders etc.), the last phase is recommended to be called "implementation".

Having understood this process, risks can be divided by the development phases and analysed independently.

4.1 Preprocessing

In the preprocessing phase, the required data is selected and transformed. Generally, this phase contains the following items based on Raschka (2015):

- feature extraction
- feature scaling
- feature selection
- dimensionality reduction
- sampling
- division of the data into different datasets.

Data quality is inevitably contributing to the prediction power of the utilised solution. If we consider as an example, that the organisation intends to predict whether a customer is going to pay back its loan, then having a dataset about the company's sales will not meet management expectations. Therefore, the first threat that should be addressed is whether the data is relevant for the required purposes, and if not, an operational risk is in place. Let's suppose there is an interrelation between the data and the target. Then what can possibly go wrong? Ribeiro et al. (2016) performed a research work regarding a binary classifier that could distinguish huskies and wolves on selected pictures. The algorithm performed relatively well, however, after deeper analysis, it has been revealed that each of the pictures that showed a wolf, there was snow at the back. If a husky was seen on the picture with a snowy background, then the algorithm automatically classified it as a wolf. That leads to two conclusions. First, the system has to generalise well on unseen data in order that it could be implemented into production environment meaning that the input dataset must be complete and accurate that might be the hardest task to achieve as for many tasks, especially for complex business problems, the universe of possible inputs determining the outcome is nearly infinite. Second of all, as Ng (2018) highlights, the input data used in production shall be similar as the training data on what the algorithm learnt the interrelations. Ng (2018) demonstrates the problem with a cat recognition application. If there is an algorithm that learns from data that were obtained through the internet, but performs prediction on data that was uploaded by users recorded by their mobile phone, the algorithm is not going to have a high accuracy as the training data quality differs from production data. The same problem occurs if the data is outdated, e.g., collected from the past, but predicting for the future with changed business conditions, data has missing values which has to be corrected or eliminated (Hastie et al., 2009), or the data contains outliers.

Semantical problems can also occur that is due to inappropriate human interpretation that might be one of the hardest issues that can emerge, as human error always counts to be critical in any task to be performed. Linden et al. (2019) highlights as a key challenge, that the quantity of the data is also questionable in many cases as engineers usually get to know whether the algorithms had enough data to capture the interrelation, once the problem is solved, therefore the sufficient amount of data puts a high risk for the development schedule affecting project budget. If data is not enough, then data acquisition or other data gathering techniques shall be in place that may also increase costs. There is a phenomenon called 'curse of dimensionality' that not only affects intelligent system development, but also other fields of engineering. It was first introduced by Bellman (1961) and it says that the number of samples that are needed for a problem to be solved (in our case a classification) with a determined level of accuracy increases exponentially with respect to the number of input variables. This means that if the number of data attributes is e.g., 10, then the number of data points needed are 2 power 10, i.e., 1024 and so on, meaning that the more attributes the dataset has, more datapoints are needed leading to the same conclusion, thus the budget for data acquisition increases.

Several algorithms are performing better once the input data were transformed. Deriving the lessons learnt from the publication of Rashid (2016), neural networks are performing better if the data are standardised or normalised. That is a similar case to clustering algorithms, the results are better if the data are in the same scale as variables with higher values can dominate the calculation, thus distorts the outcomes. This distortion can also be observed when there are highly correlated variables among the data meaning if one datapoint changes, then other changes, respectively, therefore it makes it harder to discover the explaining variables. However, algorithms, such as neural networks, can neglect this problem as it automatically reduces corresponding weights (gives less importance) for correlated variables. The issue can be solved by dimensionality reduction techniques such as principal component analysis (Sajtos and Mitev, 2007) for other more sensitive algorithms. In addition, ever since the General Data Protection Regulation came to force, special attention is needed to appropriately transfer data to make them not contain any personal information, one of the techniques can be anonymisation to solve this issue (Barta, 2018a).

The last item on the list in the preprocessing phase is dividing the data into training and testing datasets. It can be read in several research work e.g., in the work of Kása (2011) that even the whole dataset shall be divided into three categories, the last one is the validation set. The training set is the data on what the training is performed, the validation set is the data on what the hyperparameter tuning is taking place i.e., configuring the algorithm to reach desired accuracy, and the testing set is the dataset on what the system is evaluated. There should be an appropriate balance to be found. If the dataset is not divided and learning is performed on the whole dataset, filtering out the noise is impossible, and the system may learn interrelations that only occurred in specific cases. Another issue with this, that evaluation cannot be performed independently that may mislead performance KPIs. At this very moment, there is no optimal parameter determined regarding this issue, there are only recommendations such as published by Ng (2018), that generally the following distribution can be observed for the training set, validation set and testing set: 70%, 10%, 20%. This might be different in specific business domains.

After the examination of the aforementioned cases, it can be concluded that for the preprocessing phase the intelligent system is vulnerable against misinterpreted and erroneous data, and data shall be cleaned and quality must be provided, before any

programming activity is actually taking place, since it has an impact of the whole prediction power of the application. IT risk assessment shall be carefully planned to address this risk as consequently whichever intelligent algorithm is going to be further used, none of them will tolerate incorrect data, and wrong conclusions might be derived if used.

4.2 Learning

In the learning phase, the training data is loaded to the algorithm and the algorithm is ready to process it (after the careful selection of the algorithm with modelling techniques). Based on the interrelation found in the processed data the system is deducing a conclusion. The conclusion is determined by the internal rules that the system has declared based on the training data, but these rules are not always to be interpreted by human beings. As mentioned earlier e.g., neural network is a type of algorithm that is operating in a high dimensional space that basically cannot be understood why the conclusion is what it is. That phenomenon is called the black-box effect. The user can be ascertained that the algorithm works e.g., from a picture, it is able to identify a specific object, but the how is questionable. That can lead several problems. In the research work of Wang and Kosinski (2018) a classification task was to determine of a face of a human being of their sexual orientation. The algorithm has achieved 91% accuracy in case of men, and 83% in case of women, while repeating the same task with humans, the results were 61% and 54%, respectively. This means that the algorithm has found interrelations that even a human cannot understand, even the human results were only close to a guess like when flipping a coin and then you guess it would be heads or tails. This means that in case of the use of black-box algorithms, one can never be sure the nature and extent of interrelations what the data contains. The algorithm can discriminate people, violate regulations in order to achieve optimised performance or result in other unethical decisions. A similar case can be studied in the research work of Wilson et al. (2019) where the authors noted that the skin colour can have an impact on whether a self-driving car will hit the pedestrian or not. Therefore, when performing a risk assessment on systems utilising black-box algorithms a threat shall be addressed whether the algorithm is not causing reputational damage to the company by making unethical, unlawful and unacceptable decisions that can only be assured if the system has gone through a comprehensive testing, analysing each and every considerable test scenario and the whys are further analysed. Typical black box algorithms are neural networks (e.g., deep learning), support vector machines and random forests (Chen et al., 2005; Zhang and Zulkernie, 2016).

Additional problems can occur in the learning process due to coding problems or the utilised models are inappropriate to solve a particular task. Neural networks are not always the best solutions to use, especially, when there is not sufficient data or there is enough priori knowledge on a specific problem so that basic modelling may be sufficient.

4.3 Evaluation

In the evaluation phase the prediction power of the systems is measured among several predefined KPIs that often contains the accuracy, logarithmic loss, cross validation results and the analysis of confusion matrices. It is inspected whether the algorithm was able to learn from the data or otherwise it may be making random decisions. In addition, the data may be full of random noise indicating that the algorithm might learn this noise and therefore it will not generalise well, i.e., memorised the training set, but

underperforms on the test set or in worst case, in production. The first event is called underfitting, the latter one is overfitting. Underfitting means that the algorithm has a low accuracy on the training set (high bias) and test set, respectively, and based on the work of Ng (2018) either more data shall be collected, or a more complex algorithm shall be used. Overfitting indicates that a high accuracy was achieved on the training set, but low accuracy on the test set (high variance) that can be corrected by either regularising the data (assigning penalty values to high weights) or using less complex algorithms, such as applying the rules from Occam's razor (Pitlik, 2014). Experiences show that usually having a low bias and low variance at the same is quite hard, when the bias is decreasing than variance is increasing and vice versa (Ng, 2018).

Having appropriately predefined evaluation metrics is a key, and risk shall be assessed whether the particular problem to be solved is within the defined criteria and adequate indicators were used to measure the effectiveness. For instance, in case of credit card fraud detection, most probably the dataset on what the algorithm was working on, contains only just a few examples that are fraudulent that might be even under 1% of the whole dataset. In this case the algorithm will by default achieve a high accuracy (by evaluating corresponding error matrix), at least 99%, that can be said good in case of many other applications, however, in this case, the algorithm may not have been able to classify fraudulent transactions, and guessed that each and every transaction was correct.

4.4 Prediction

The prediction phase is when the system is ready to be deployed into the production environment. Developers, dedicated testers and business users also tested the application, and management feels satisfied with the results and produced accuracy. There are general IT risks and specific risks too, to be addressed. Appropriate user access management is considerably important at this phase, as developers cannot perform the implementation. The reason is that once business units are done with the testing and the system is approved, no opportunities shall be provided to modify system configuration, data, program code etc. as it might negatively affect the application and its prediction power. This prevents unauthorised access. However, further development might be needed, but then the whole testing process must be repeated from the very beginning. Clark (2018) highlights that the system may communicate with other applications, thus interfaces and the whole IT environment must be tested. Other thing to consider is that system development did most probably not stop at this point, as the data are collected and fed into the algorithm to further improve its accuracy in time, meaning that continuous monitoring and testing is inevitable, thus the risk assessment shall also address risk such as having and educating appropriate staff that are dedicated to perform this job on a daily basis, and as mentioned before, the quality of data, evaluation of the system etc. In addition, there is another phase in traditional system development that is the post-implementation review that involves business analysis that must be performed to verify that the investment was beneficial, i.e., there is more benefit and profit on a long run than cost, established KPIs are met, and that the system is operating as management intended. Another security risk arises if the deployment is made in the cloud in case if it is e.g., outsourced to a service provider. In this case, general IT risk assessment must be performed to assess system exposures and contract risk compliance shall be conducted.

5 Frameworks for AI risk management

The paper has until this point analysed the direct risks arising from the development of AI, however, implementing AI based solutions also has risks on the business as a whole, therefore AI applications shall be analysed from business points of views, respectively. The Committee of Sponsoring Organizations of the Treadway Commission (COSO) Enterprise Risk Management Framework (2017) can supply with appropriate information that can help incorporate AI risks into the company's ERM. AI risk assessment shall be an integral part of the enterprise risk management framework that addresses other risks, such as regulatory, financial, credit, IT etc. If there is an organisational risk management framework implemented throughout the whole enterprise, it can assure that risk assessment procedures are not ad hoc, i.e., there is program that is continuous and enforced with management approved policies. The framework focusing on ERM has five interrelated components:

- governance and culture
- strategy and objective-setting
- performance
- review and revision
- information, communication, and reporting.

These five components can guide companies through the management of internal risks. Based on the publication of the Federation of European Risk Management Associations (FERMA) (2019), so as to manage AI risks, companies shall perform an analysis to determine their risk profile, and the COSO ERM (2017) can give an appropriate foundation for such an analysis. The FERMA has published an AI risk management framework that is shown in Table 1.

Table 1 Scope of AI-relates risks

Strategic and environmental risks	*Business risks*	*Operational risks*
Governance	Product conception	IT risks: data and information
Data and infrastructure governance	Product production	Organisation and project management
Liability	Product distribution	Human resources
Environmental impact	Market disruption	Externalisation
Data property/sovereignty breach		Continuity of activity/recovery plan

Source: Federation of European Risk Management Associations (2019)

Based on the analysis of the framework, in respect to AI applications, there should be a companywide strategy developed for the implementation of AI based business process automation that means that the goals and objectives to be achieved are appropriately planned, aligned with business goals and follows the strategic directions of the organisation. As it was revealed from the study of Andrews et al. (2017) the second greatest challenge that companies are facing nowadays when it comes to AI implementation is the definition of their AI strategy. Without adequately harmonised AI strategy, there is a risk that the business cannot benefit from AI implementation that can

cost lots of used resources and capital. When defining strategy, companies shall consider regulatory requirements and other relevant compliance areas. The personnel being accountable and responsible for overseeing the programs shall be defined, policies, procedures, standards and guidelines shall be developed in order to establish criteria and enforce a healthy, compliant and management approved business environment. The framework also addresses business risks regarding product conception, production and distribution, and market disruption. This indicates that the product that has AI capabilities built in, shall be prudently designed and developed, in addition, disruption arising from unintended consequences shall be eliminated. The third risk category is the operational risks. Operational risks are considered for supporting processes when AI is utilised to automate these processes, however, AI may cause headache if risks are not managed properly. These processes can include, depending on the nature and extent of the utilised AI solutions, organisation and project management procedures, HR, externalisation (third party risks), business continuity issues etc.

Another framework was developed by Deloitte, a strategic framework for algorithmic risk management that can be used as an approach to effectively manage risk arising from AI based process automation. Table 2 summarises the framework that is now understood as a tool to coordinate the risk program for intelligent systems.

Table 2 Framework for algorithmic risk management

Strategy and governance		*Design, development, deployment, and use*	*Monitoring and testing*
Goals and strategy	Principles, policies, standards and guidelines	Algorithm design process	Algorithm testing
Accountability and responsibilities	Life cycle and change management	Data assessment	Output logging and analysis
Regulatory compliance	Hiring and training of personnel	Assumptions and limitations	Sensitivity analysis
Disclosure to user and stakeholder	Inquiry and complaint procedures	Embedding security and operations controls	Ongoing monitoring
Inventory and risk classification		Deployment process	Continous improvement
		Algorithm use	Independent validation
Enterprise risk management			

Source: Krishna et al. (2017)

As similar to the previously presented framework, the strategy and governance are also indicated as the basis of AI implementation and risk management. One important to note difference is that in the previous framework, environmental impact is additionally considered. Strubell et al. (2019) also pointed out that training neural network models for natural language processing purposes (NAS model) emit carbon dioxide (approximately 284 ton), slightly less than altogether five average American car in their whole lifetime, therefore environmental impact is also a valid point for debate.

The Deloitte framework includes another risk category that the "Design, development, deployment, and use". This element focuses on technical processes that ensures that system design, development, deployment and use are controlled and reviewed in alignment with good practices and internal procedures. System design shall be in place to analyse whether the new system contains the functions that the business expects, technically can be implemented in the IT environment and can communicate with other systems (e.g., data transfer), and also IT security is considered. System development shall lie on project management principles, shall be continuously tested, bugs shall be fixed in a timely manner. Before the deployment, testing must be in place to verify that business users are satisfied with the results and deployment is performed to production environment by independent functional users preventing any not tested or approved parts of the system to be used in production. The last element is the "Monitoring and testing". The system needs continuous performance monitoring to ensure that the system is functioning as business intended and the system is maintained to be adjusted for future use even in case of changes in the business environment. In addition, past performance is not always a reliable indicator of future performance of the system, thus, monitoring its accuracy and other KPIs are essential. This framework, therefore, is more focusing on the risk management of the AI development lifecycle, unlike the previous one had the business processes in centre.

A third framework herein introduced is developed by KPMG (2018) and is shown in Table 3.

Table 3 AI risk and controls matrix

Strategy					
Governance					
Human resource management	Supplier management	Risk management and compliance	Enterprise architecture	Data and model governance	Knowledge management
Program governance and management		Business process controls			
		Logging and monitoring			
Solution development		Security Management	Identity and access management		
		IT change management	IT operations	Business continuity	

Source: KPMG (2018)

The framework contains 17 categories for managing risks and controls for AI systems and is referenced to COBIT processes and areas. This is the one framework from all introduced three that is also giving example controls to cover the emerged risks arising from the implementation of AI. The framework covers the general business and IT processes regarding the development and implementation of AI applications, however, it does not address specific requirements, that are unique for AI development. This contains the following items, but not limited to:

- *Identity and access management*: Since most of the systems, and especially AI, operate with tons of confidential data, it is essential to assess who can directly access these pieces of information. If someone has access to data that are fed with learning algorithms, then this person might be able to manipulate i.e., delete, modify or create new or old inputs being able to totally mislead the system e.g., deceive it in case it would show better results for quality assurance, or the system would not be able to detect fraudulent activities.
- *Change management*: Systems have security or other configurations, parameters etc. that is to maintain system resistance or just needed for system setup to operate in the company's environment as expected. If changes are not recorded, logged, backed up, then misconfiguration can lead to malfunction. Change management controls shall be assessed whether they are in place to test, approve and store any configuration changes that makes this process transparent for its users and the management. This also includes underlying databases or operating platforms.
- *Business continuity*: Systems may experience malfunction in the production environment after deployment, and therefore, the old versions may need to be restored. This requires rollback planning, and tested business continuity processes.
- *Logging and monitoring*: One of the most important controls that shall exist is the analysis of audit logs. Audit logs are the primary source for every malfunction that is experienced with IT systems, therefore they must be detailed and no writing access should be granted in order to maintain its integrity. Audit logs must provide information that who, when, where accessed the systems, and what was created, changed, deleted etc.

In addition to general IT risk concerns, legislative and regulative rules must be followed and adhered to. In many cases e.g., face recognition applications, data subjects must contribute their consent for the data processing activity that means that regulations such as GDPR must be considered from the very beginning of the system development.

An interesting difference is that, this is the framework where the 'expectation mismatch' is discussed. That means that business leaders are sometimes not aware of the feasibility of the wanted AI solution, they want something that AI cannot do as of today, and thus, they are disappointed in the results. Appropriately skilled consultants and experts shall be involved in the development and implementation in order to avoid the expectation mismatch.

After reviewing the three frameworks, the opinion of the authors is that companies shall adopt as soon as possible an AI related risk assessment model into their ERM in order to get prepared for the risks that AI development and implementation contain. The introduced frameworks complete each other, however, none of them are talking about the following risks that should be added to a hybrid framework of the three:

- *Lack of computer power*. AI algorithms, especially the Deep Learning, are computationally expensive, thus without the appropriate hardware resource, training these models can take days, weeks or so. There is also an option to outsource the process, however, a cost-benefit analysis should be in place to determine which model fits the best to the company's profile.

- *Impact on clients*. AI can have a great impact on service delivery, therefore risks shall be identified that may have a negative impact on clients. An example can be a customer service chat bot that is not developed or implemented appropriately to completely satisfy customer queries.

With the use of an enterprise wide risk assessment framework, organisational and technical controls can also be developed to establish a controlled environment for system development. Definitely, the frameworks detail only high level processes, however, it is important to keep in mind, that different problems require different approaches, or even processes, therefore not only policy level regulation shall be in place, but operating procedures must be followed that further details the objectives of the policies on process or project level.

6 Mitigating/compensating controls

The result of a comprehensive IT risk assessment is the preparation of the risk treatment plan. The document regularly includes the affected IT assets and risks that shall be treated by a management approved strategy. If the organisation does not accept, avoid or transfer the risk, then they shall mitigate them by implementing controls. Risks shall have risk owners that are in charge of keeping track of the risk monitoring process and have the authority to implement risk reducing countermeasures. Without an owner, the risk will not be reduced as there are no responsible personnel appointed. In addition, the severity of each risk shall be determined in order to prioritise tasks and risks to be handled and so then projects can be built to compensate them. Definitely, if every countermeasure is implemented to eliminate the determined risks, the IT risk assessment still should be continuous as new threats and vulnerabilities arise due to changes in business, operation, legislative, technological etc. environments and therefore they need to be assessed and managed. Based on the analysis of what the authors performed the following illustrative risks, as a summary, shall be assessed and compensated by implemented controls (in alignment with IT risk assessment results):

- *Data management*: Data for AI application shall be managed. Processing integrity and access security shall be in place.
- *Application development*: Learning algorithm shall be carefully chosen for particular problems. Testing, approval, deployment procedures shall be in place. Programs shall be debugged and free of programming failures.
- *Evaluation criteria*: Metrics shall be developed closed to the problem to be solved that expresses the main objectives. Preferred to be developed in advance.
- *General IT risks*: General IT risks shall be considered that means the assessment of access security to every system component, change management, recovery procedures, audit log analysis, encryption, incident management etc.
- *Regulations*: Local and global regulations and internal policies shall be followed; internal strategy shall be developed in alignment with business objectives.

7 Conclusions

The main objective of the paper was to understand the general main risks for AI based business process automation applications that shall be considered when performing IT Risk assessments. The authors performed literature reviews to obtain an overview of the procedures and steps to be performed for conducting general IT risk assessments and the development stages of AI applications, that has led to the following conclusions:

- General IT risk assessment processes can be utilised (such as ISO 27005), since AI systems can be deemed to be IT assets similarly to traditional software, however, intelligent systems require customised threat catalogues and compensating controls, thus risk assessment methodologies shall be adjusted to specific needs of such systems. This statement is supported by several examples that was processed as part of the research work.
- Before the development, data management and data quality controls are essential to be implemented so as to ensure that the system is processing data that can mostly contribute to the best achievable prediction power including the elimination of any semantical pitfall. The risk assessment must deeply consider data manipulation threats, since algorithms used in intelligent decision making are highly vulnerable against not appropriately transformed and selected data sources.
- The use of black-box algorithms, in worst case scenario, can lead to the loss of good business reputation, since this kind of algorithms cannot be interpreted in terms of concluding logic. This means that these algorithms may make decisions that are not ethical or lawful. The risk program shall address whether to use black-box algorithms for a particular business problem and if so, data used to feed the system was gathered in alignment with external regulations and with the consent of data subjects. When black-box algorithms are to be used, the authors recommend its usage to be limited for decision support, not decision making.
- Evaluation metrics used to performance measurement shall be developed problem specifically, since even the distribution of the different categories of data may require different KPIs. The risk assessment program shall consider whether appropriate metrics are in use for measuring system performance.
- Intelligent systems shall be continuously maintained as the operating environment might change, therefore new obtained data have to be processed and systems need continuous testing to ensure integrity. In addition, general IT risks also shall be considered as they are still applicable for intelligent systems. Using a business process-based framework may seem to be a great decision to establish the boundaries of system development and a controlled environment that enforces technical and organisational controls, respectively.

The research work performed were examining the general aspects of intelligent automation, and were not investigating specific algorithms in details. The paper presented an overview of recent AI failures and risks that were summarised along with system development that may help for further researchers to see AI risks as a whole on a system level.

The authors recommend the development of AI specific risk assessment methodologies and standards including customised threat catalogues in order that

companies can manage risks arising from the use of intelligent systems. Further research is also recommended to understand how black-box algorithms are working as these methods appear to result in high prediction power.

References

Andrews, W., Sau, M., Dekate, C., Mullen, A., Brant, K., Revang, M. and Plummer, D. (2017) *Predicts 2018: Artificial Intelligence*, Retrieved from https://www.gartner.com/document/3827163?ref=solrAll&refval=193910164&qid=780b332f7d9afba6f17865ea8b939339

Barta, G. (2018a) 'Challenges in the compliance with the general data protection regulation: anonymization of personal information and related information security concerns', in Paweł, U. and Paweł, W. (Eds.): *Knowledge – Economy – Society Business, Finance and Technology as Protection and Support for Society*, pp.115–121, Foundation of the Cracow University of Economics, Poland.

Barta, G. (2018b) 'The increasing role of IT auditors in financial audit: risks and intelligent answers', *Business Management and Education*, Vol. 16, No. 1, https://doi.org/10.3846/bme.2018.2142, pp.81–93.

Barta, G. and Görcsi, G. (2018) 'Artificial intelligence and audit: why is it necessary to audit the intelligent decision support?', in Földi, P., Borbély, A., Kápolnai, Z., Zsarnóczky, M.B., Bálint, C., Fodor-Borsos, E., Gerencsér, I., Gódor, A.K., Gubacsi, F., Nyírő, A. and Szeberényi, A. (Eds.): *Közgazdász Doktoranduszok És Kutatók, I.V. Téli Konferenciája*, Doktoranduszok Országos Szövetsége, Hungary, pp.225–234..

Bellman, R.E. (1961) *Adaptive Control Processes*, Princeton University Press, Princeton, NJ.

Borgulya, I. (1998) *Neurális Hálók És Fuzzy-Rendszerek*, Dialóg Campus Szakkönyvek, Budapest – Pécs.

Chen, W.H., Hsu, S.H. and Shen, H.P. (2005):'Application of SCM and ANN for intrusion detection', *Computers and Operations Research*., Vol. 32, No. 1, pp.2617–2634.

Chollet, F. (2018) *Deep Learning with Python*, Manning Publications Co., New York.

Clark, A. (2016) *Focusing IT Audit on Machine Learning Algorithms*, Retrieved from https://misti.com/internal-audit-insights/focusing-it-audit-on-machine-learningalgorithms

Clark, A. (2017) *Machine Learning Audit in the 'Big Data Age'*, Retrieved from https://www.cioinsight.com/it-management/innovation/machine-learning-audits-in-the-big-data-age.html?lipi=urn%3Ali%3Apage%3Ad_flagship.3_profile_view_base%3BZBLUSmhrSLqwbpJa%2F%2BH7Wg%3D%3D

Clark, A. (2018) *The Machine Learning Audit–CRISP-DM Framework*, Retrieved from https://www.isaca.org/Journal/archives/2018/Volume-1/Pages/the-machine-learning-audit-crisp-dm-framework.aspx?lipi=urn

Committee of Sponsoring Organizations of the Treadway Commission (2017) *Enterprise Risk Management. Integrating with Strategy and Performance*, Executive summary. Retrieved from https://www.coso.org/Documents/2017-COSO-ERM-Integrating-with-Strategy-and-Performance-Executive-Summary.pdf

Dua, S. and Du, X. (2011) *Data Mining and Machine Learning in Cybersecurity*, Taylor and Francis Group, USA.

Federation of European Risk Management Associations (2019) *Artificial Intelligence Applied to Risk Management*, Retrieved from https://www.eciia.eu/wp-content/uploads/2019/11/FERMA-AI-applied-to-RM-FINAL.pdf

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. 2nd ed., Springer Science, New York.

Information Security Forum (2014): 'IRAM2.The Next Generation of Assessing Information Risk. Information Security Forum Limited.

ISACA (2009) *The Risk IT Framework*, ISACA, USA.

ISO/IEC (2011) *27005 Information Technology – Security Techniques – Information Security Risk Management*, International Standard, Switzerland.

Kása, R. (2011) *Neurális Fuzzy Rendszerek Alkalmazása a Társadalomtudományi Kutatásban Az Innovációs Potenciál Mérésére*, Doktori disszertáció, Retrieved from http://193.6.1.94: 9080/ JaDoX_Portlets/documents/document_6323_section_1701.pdf

KPMG (2017) *AI Risk and Control Matrix*, Retrieved from https://assets.kpmg/content/dam/ kpmg/uk/pdf/2018/09/ai-risk-and-controls-matrix.pdf

Krishna, D., Albinson, N. and Chu, Y. (2017) *Managing Algorithmic Risks. Safeguarding the Use of Complex Algorithms and Machine Learning*, Retrieved from https://www2.deloitte.com/ content/dam/Deloitte/us/Documents/risk/us-risk-algorithmic-machine-learning-risk-management.pdf

Linden, A., Mullen, A. and Khan, A. (2019) *Boost Your Training Data for Better Machine Learning*, Retrieved from https://www.gartner.com/document/3953637?ref=solrAll & refval=229934751
& qid=ed1c5b1eea5ba5a099fd8b0

McCarthy, J., Minsky, M., Rochester, N. and Shannon, C.E. (1955) *Proposal for the Dartmouth Summer Research Project on Artificial intelligence*, Tech. Rep.3p., Dartmouth.

MMC Ventures (2019) *The State of AI 2019: Divergence*, Retained from https://www.mmcventures.
com/wp-content/uploads/2019/02/The-state-of-aI-2019-divergence.pdf

National Institute of Standards, and Technology (2012) *Guide for Conducting Risk Assessments*, NIST Special Publication 800-30, USA.

Ng, A. (2018) *Machine Learning Yearning. Technical Strategy for AI Engineers' in the Era of Deep Learning* (Draft ver.), deeplearning.ai, USA.

PCI Security Standards Council (2012) *Information Supplement: PCI DSS Risk Assessment Guidelines*, PCI Data Security Standard (PCI DSS).

Pitlik, L. (2014) *Occam Hermeneutikája. Magyar Internetes Agrárinformatikai Újság*, Retrieved from http://miau.gau.hu/miau2009/index.php.3?x=e0& string=occam

Raschka, S. (2015) *Python Machine Learning*, Packt Publishing, Birmingham.

Rashid, T. (2016) *Make Your Own Neural Network*, 1st ed., CreateSpace Independent Publishing Platform.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 'Why should I trust you?', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'16*, San Francisco, California, USA.

Russel, J. (2017) Google'*s AlphaGo AI Wins Three-Match Series Against the World's Best Go Player*, Retrieved from https://techcrunch.com/2017/05/24/alphago-beats-planets-best-human-go-player-ke-jie/

Russel, S. and Norvig, P. (2005) *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson Education, India.

Sajtos, L. and Mitev, A. (2007) *SPSS Kutatási És Adatelemzési Kézikönyv*, Alinea Kiadó, Budapest.

Schuett, J. (2019) *A Legal Definition of AI*, Retrieved from https://arxiv.org/pdf/1909.01095.pdf

Strubell, E., Ganesh, A. and McCallum, A. (2019) *Energy and Policy Considerations for Deep Learning in NLP*, Retrieved from https://arxiv.org/pdf/1906.02243.pdf

Wang, Y. and Kosinski, M. (2018) 'Deep neural networks are more accurate than humans at detecting sexual orientation from facial images', *Journal of Personality and Social Psychology*, Vol. 114, No. 2, pp.246–257.

Wilson, B., Hoffman, J. and Morgenstern, J. (2019) *Predictive Inequity in Object Detection*, Retrieved from https://arxiv.org/pdf/1902.11097.pdf

Yampolskiy, R.V. and Spellchecker, M.S. (2016) *Artificial Intelligence Safety and Cybersecurity: A Time of AI Failures*, Retrieved from https://arxiv.org/ftp/arxiv/papers/1610/1610.07997.pdf

Zhang, J. and Zulkernie, M. (2016) 'A hybrid network intrusion detection technique using random forests', *Proceedings of the First International Conference on Availability, Reliability and Security*, pp.262–269.