No AI risk if you don't use AI? Think again!

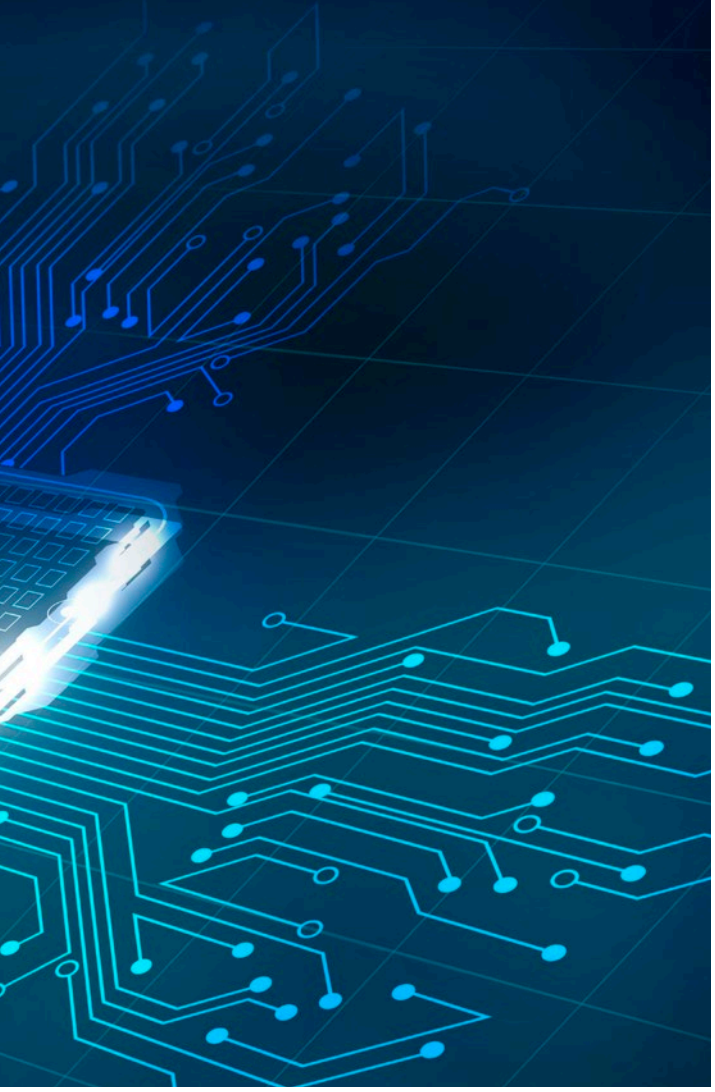
Dr. Alexander Boer
is a senior manager at KPMG Trusted Analytics.
boer.alexander@kpmg.nl

AI-related risks regularly make news headlines and have led to a number of legislative initiatives in the areas of privacy, fair and equal treatment, and fair competition. This may cause organizations to shy away from using AI technology. AI risks, as commonly understood, are however caused largely by the degree of autonomy and the increasing social impact of data processing rather than just by new algorithms. These risks should be understood holistically as threats to entire IT infrastructures rather than to individual AI components. A broad, comprehensive and ongoing AI-related risk assessment process is essential for any organization that wants to be ready for the future.

INTRODUCTION

Computers don't always do what you want, but they do what they were instructed to do. This clearly separates the computer as an agent performing a task from a human being doing the same. Computers as components in a business process are in essence predictable: their behavior follows a design specification, and the same input will generate the same output. People, on the other hand, are the unpredictable components of a business process. In practice, they often do not fully follow instructions. They deviate from the business process specification, for bad and for good reasons. People are autonomous.

On the one hand, people are a weak point and therefore form a major risk. They may be sloppy, slow, commit frauds, extract confidential data for their own purposes, be influenced by unconscious biases, etc. On the other hand, people often take the rough edges out of a business process. People use their own common sense, see new patterns in data, spontaneously remedy injustices they see, diagnose problems in the business process, are aware of changes in society that may affect business because they follow the news, and generally generate valuable feedback for adapting and continually improving business processes. People make processes more resilient.



The following examples of news headlines are typical of the stories that may be attributed to AI-related risk, in the sense that algorithm-based decision-making is at the center of the story, although they deal with completely different forms of error:

- Google accused of secret program giving them an unfair advantage in ad-buying ([Feis21])
- Amazon sells a $23,698,655.93 book about flies ([Eise11])
- Court rules Deliveroo used discriminatory algorithm to assess its riders ([Geig21])
- MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs ([Quac20])
- There's software used across the country to predict future criminals and it's biased against blacks ([Angw16])

Complexity essentially deals with how easy it is to simulate the behavior of a system in our mind, on the level of abstraction we care about. What we require of this simulation largely depends on our needs for explainability. For instance, a facial recognition application is, objectively from an information theoretic perspective, more complex than a simple risk-scoring model based on social-economic parameters. Since we usually do not wonder how we recognize faces, we tend to take its behavior on a functional level for granted, until we discover it makes mistakes we would not make. Only then we face a complexity problem.

BLACKBOXNESS

Popularly, AI technology is positioned somewhere between humans and computers. It has, in essence, a *blackboxness* problem. It may have some capacity to adapt to changes in its environment. It sometimes surprises us by finding predictive patterns in data we did not see. But its design specification does not lend itself to simulation of its behavior in our mind: the relation between input and output data is discovered by the AI technology itself. It is not predictable. Not to us. And it does make mistakes that humans will never make. Mistakes that are hard to explain. Sometimes the mistakes are even hard to notice.

Because *blackboxness* is bad English we will call it a complexity problem instead, keeping in mind that we do not have an objective measure of topological complexity in mind, but rather our inability to simulate what it does. AI technology is, therefore, *complex*.

AI-related risks regularly make news headlines, may cause significant reputation damage, and have led to a number of legislative initiatives and ethical frameworks in the areas of privacy, fair and equal treatment, and fair competition. The associated cost of introducing effective control measures may cause organizations to shy away from using AI technology, or to pick traditional, well-established techniques for data analysis in favor of more complex and more experimental ones. We see a preference for linear regression techniques in many organizations for exactly this reason. This is not a solution. While shying away from AI technology may be a valid choice in certain circumstances, it neither addresses the inherent risks nor necessarily exempts one from special legal responsibilities.

In this article we address the origin of some of the inherent risks, and the role AI and data play in these risks, and finally come to the conclusion that a broad, comprehensive and ongoing AI-related risk assessment process is essential for any organization that wants to be ready for the future.

IS IT AI?

A first problem becomes apparent if we look at European legislative initiatives that create potentially expensive compliance requirements. There is no overarching agreement about the kinds of systems that create AI-related risks. This is not surprising because the risks are diverse and take many forms.

Let us quickly run by some examples. Art. 22 of the GDPR is already in effect and targets automated decision making using personal data – regardless of the technology used. Besides the limitations to personal data, there is a clear concern regarding the degree of autonomy of systems. The freshly proposed Artificial Intelligence Act ([Euro21a]) prohibits and regulates certain functions of AI based on risk categories – instead of starting from a restrictive definition of technology. For a civil liability insurance regime for AI ([Euro20]) it is too early to tell how it will turn out to work, but it makes sense that it will adopt a classification by function as well.

The Ethics Guidelines for Trustworthy AI ([Euro19]) on the other hand target technology with a certain adaptive – learning – capacity, without direct reference to a risk-based classification based on function. This is a restrictive technology-based definition, but one which leaves big grey areas for those who try to apply it. The Dutch proposed guideline for Government agencies ([Rijk19]) targets data-driven applications, without a functional classification, and without reference to learning capacity.

This already creates a complicated scoping problem as organizations need to determine which classifications apply to them and which do not. And beyond that there is legislation that directly impacts AI but does not directly addresses it as a topic. Existing restrictions on financial risk modeling in the financial sector obviously impacts AI applications that make financial predictions, regardless of the technology used. New restrictions on self-preferencing ([Euro21b]) will for instance impact the use of active learning technology in recommender algorithms but they will be technology-agnostic in their approach.

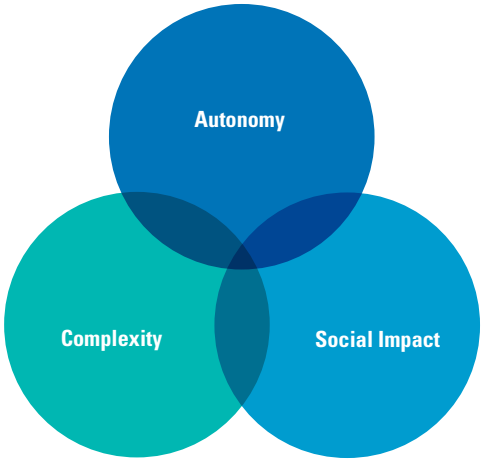


Figure 1. Three dimensions of AI risk.

AI risk may address software that you already use and never classified as AI. It may address descriptive analytics that you use for policy making that you never considered as software, and don't have a registration for. Your first task is therefore to review what is present in the organization and whether and how it is impacted by compliance requirements related to AI. Beyond that, seemingly conflicting compliance requirements will create interpretation problems and ethical dilemmas. For instance, when you have to choose between privacy protections of the GDPR on the one hand and measurable non-discrimination as suggested by the Artificial Intelligence Act on the other, and both cannot be fully honored.

THREE DIMENSIONS OF AI RISK

All in all, we can plot the risk profile of AI technology on three different dimensions. Although the risks take diverse forms, the underlying dimensions are usually clear. The first one is the one we already identified as complexity.

But complexity is not the major source of risk. AI risk is predominantly caused by the degree of autonomy and the increasing social impact of data processing rather than just by new algorithms. Risks are often grounded in the task to be performed, regardless of whether it is automated or not. If how well the task is executed matters significantly to stakeholders, then risk always exists. This is the risk based on its social impact. If the automated system functions without effective oversight of human operators, it is autonomous. Autonomy is the third source of risk. We also regard it matter-of-factly autonomous if human operators are not able to perform the function of the system, either because they cannot come to the same output based on the available input data, or because they cannot do so within a reasonable time frame.

If an automated system scores on any of these three dimensions (see Figure 1), it may carry AI-related risk with it if we look at it within its data ecosystem. This is not because one single dimension creates the risk, but because a source of risk on a second risk dimension may be found nearby in the IT infrastructure, and we need to check that.

DATA ECOSYSTEMS

Most AI-related risks may also surface in traditional technology as decision making ecosystems are increasingly automated. Increasing dependence on automation within whole task chains causes human decision mak-

ers to increasingly go out of the loop, and the decision points at which problems could be noted by human decision makers in the task chain become few and far between. The risks are caused by the increasing autonomy of automated decision-making systems as human oversight is being reduced. If things go wrong, they may really go wrong.

These risks should be understood holistically as threats to entire IT infrastructures rather than individual AI components. We can take any task as our starting point (see Figure 2). When determining risk there are basically three directions to search for risk factors that we need to take into account.

Upstream task dependencies

If the task uses information produced by AI technology, it is essential to gain insight into the value of the information produced by the technology and the resilience of that information source, and to take precautions if needed. The AI technology on which you depend need not be a part of your IT infrastructure. If you depend on a spam filter for instance, you risk losing important emails and you need to consider precautions.

Downstream task dependencies

If a task shares information with AI technology downstream it is essential to understand all direct and indirect outcomes of that information sharing. Moreover, you may take specific risks, such as reidentification of anonymized information, or inductive bias that may develop downstream from misunderstanding the data you create, and you may be responsible for that risk.

Ecological task interdependencies

If you both take information from and share information to an AI component, fielding a simple task agent may increase your risks of being harmed by the AI component's failure or be exploited by it. You should take strict precautions for misbehaviors of AI components that interact in a non-cooperative setting with your IT systems. Interaction between agents through communication protocols may break down in unexpected ways.

Information usually equates with data when we think about computers, but make sure to keep an eye on information that is shared in ways other than by data. If a computer opens doors, the opening of the door is observable to third parties and carries information value about the functioning of the computer. If you open doors based on facial recognition, discrimination is going to be measurable, purely by observation.

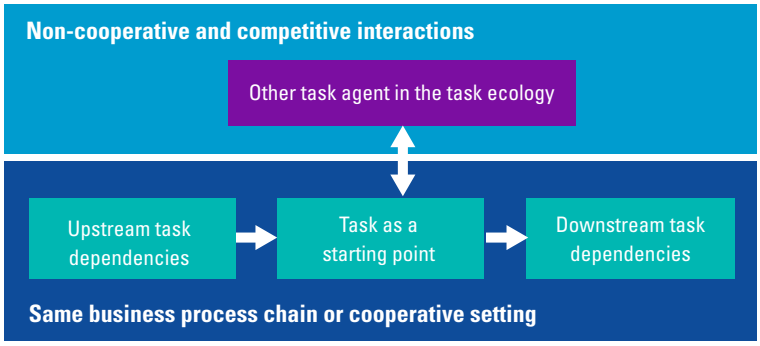


Figure 2. Where does the risk come from?

Ecologies of task agents are mainly found in infrastructures, where predictive models representing different parties as task agents function in a competitive setting. For instance, online markets and auctions for ad targeting. A systemic risk in such settings is that predictive models may cause a flash crash or collusion to limit open competition. Fielding a simple technological solution in a setting like that is usually not better than fielding a smart one from a risk point of view

DATA IS NOT HISTORY

Nearly all avoidable failures to successfully apply AI-based learning from data find their origin in either inductive bias, systematic error caused by the data you used to train or test the system, or in *underspecification*, mainly caused by not carefully thinking through what you want the system to do ([Amou20]). And besides that, there are unavoidable failures if the relationship between input data and the desired output simply does not exist. This is mainly caused by uncritical enthusiasm for AI and Big Data.

If you are a Data Scientist, it is easy to jump to the conclusion that biases in models are merely a reflection of biased ways of working in the past because historical data is used. That conclusion is, however, too simple and conflates the meaning of information and data. Not all information is stored as data, and not all data that is stored was used as information for decision-making in the past.

The information we use to make decisions is changing, and even without AI technology this creates new risks. When we remove humans from decision making, we lose information that was never turned into data. Decisions are no longer based on information gleaned from conversations in face-to-face interactions between the decision maker and stakeholders. Even if we train models on historical data we may miss patterns in information that was implicitly present when that historical decision was taken.

Inductive bias may lead to discrimination of protected groups, besides other performance-related problems. To properly label measurable inequalities ([Verm18]) as discrimination you have to understand underlying causal mechanisms and the level of control you have over those. Lack of diversity in the workplace may for instance be directly traceable to the output of the education system. As a company you can only solve that lack of diversity at the expense of your competitors on the job market.

Risk scoring models are increasingly used in government, insurance and the financial sector. The function of these models essentially works as a filter for the rule-based system, which is vulnerable to gaming the system risks because of its relative simplicity. The application of AI technology is intended to reduce the risk. KPMG Trusted Analytics has looked at risk mitigating measures taken at some government agencies to protect risk scoring models against biases. Any shortcomings we found thus far relate to the whole business process of which the risk scoring model is a part. The model itself hardly adds to the risk. Simple human-made selection rules used in those same processes were in our view considerably riskier.

Best practices regarding the privacy impact assessment (PIA) may be used as an analogy for a comprehensive AI risk assessment. In practice, many data-driven organizations have organized privacy impact assessments regarding:

- datasets,
- data-driven applications, and
- data-processing activities.

This way of working reflects an important insight about data ethics. Ethical principles about the use of personal data usually relate to either:

- reasons for collecting and storing data about people, and dealing with information about, and modification and deletion of that data,
- reasons for making such data available for use by an application, and the privacy safeguards built into that application, or
- specific purposes that such an application is put to in data-processing activities throughout the organization, and process-based privacy safeguards in those environments.

In a modern data-driven organization, the same type of data may be used and produced by various applications, and the same application may be used for different purposes in different activities. The relation between personal data and the uses to which it is put may therefore be complex and hard to trace. This complexity is managed by splitting accountability for the data between data management teams, application development teams, and business users.

Big data

At the same time, we are also tapping into fundamentally new sources of information and try to make predictions based on this. Data sharing between organizations has become more prevalent, and various kinds of data traces previously unavailable are increasingly mined for new predictive patterns. It is easy to make mistakes:

- Wrongly assuming predictive patterns are invariant over time and precisely characterize the task, and will (therefore) reliably generalize from training and testing to operational use ([Lipt18]).
- Overlooking or misinterpreting the origin of inductive biases in task dependencies, leading to an unfounded belief in predictive patterns.

Simple rules

Big Data is not just used by AI technology. Insights from Big Data may end up in automated decision-making indirectly, through new business rules formulated by policymakers based on a belief in patterns deduced from descriptive statistics on large sets of data with low information density. In essence we are doing the same thing as the machine learning algorithm, with one big difference: there is a human in the loop who confirms that the pattern is valid and may be operationally used. The statistical pattern is translated into a simple rule as part of a simple and predictable form of automation. And therefore does not carry AI risks? In reality we run the same data-related risks as before: our simple rule may turn out to be less invariant than we thought, and it may be grounded in inductive biases that we overlooked.

AI AS A MITIGATOR OF RISK

The use of AI technology instead of something else could add to the already existing risk, but it might mitigate already existing risks too. One important business case for predictive models is risk scoring, which differentiates between high risk cases and low risk cases to determine whether they may be processed automatically by a fragile rule-based system or should be handled by a human decision maker. Another important application of AI technology is detecting changes in input patterns of other systems, to make sure warning bells start ringing if a sudden change is detected. The application of AI technology *is* the risk mitigation measure in this case. It is unfortunate if these options are discarded because AI technology is perceived as too risky.

A BROAD PERSPECTIVE ON AI RISK

While AI-related compliance responsibilities may focus on the technology itself, insight in risk necessitates looking at the environment in which the technology is fielded. Few risks are inherent in the technology itself. To determine your risk profile in terms of autonomy and social impact it is necessary to look at the whole business process and its business value to the organization and other stakeholders.

Besides that, understanding data lineage is of critical importance. In a modern data-driven organization, the same type of data may be used and produced by various applications, and the same application may be used for different purposes in different activities. This complexity can be managed to some extent by clearly splitting accountability for uses of data between data management teams, application development teams, and business users.

Responsibilities for understanding the environment you work in does not stop at the boundaries of the organization however. Third-party sourcing plays a key role, just like understanding your performance in competitive settings. In certain cases, setting up network arrangements or trusted third parties for keeping control over AI risk may turn out to be a solution to preventing unnecessary duplication of work.

CONCLUSION

A broad, comprehensive and ongoing AI-related risk assessment process is essential for data-driven organizations that want to be ready for the future, regardless of whether they aim to use AI. Local absence of AI technology does not absolve you from responsibilities for AI-related risk. The big question is how to organize this ongoing risk assessment process. One element of the solution is organizing accountability for uses of data between data management teams, application development teams, and business users. Another common element of the solution may be the formation of network arrangements with other parties to reduce the cost of control. An element that is always needed, and one that the KPMG Trusted Analytics team aims to provide for its customers, is a long list of known AI-related risk factors. And another long list of associated controls that can be used to address those risks from a variety of perspectives within an organization or a network of organizations. The first step for an organization is taking the strategic decision to take a good look at what its AI-related risks are and where they come from.

References

[Amou20] D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., ... & Sculley, D. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

[Angw16] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved from: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[Eise11] Eisen, M. (2011, April 22). Amazon's $23,698,655.93 book about flies. Retrieved from: https://www.michaeleisen.org/blog/?p=358

[Euro19] European Commission (2019, April 8). Ethics guidelines for trustworthy AI. Retrieved from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[Euro20] European Parliament (2020, October 20). Recommendations to the Commission on a civil liability regime for artificial intelligence. Retrieved from: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html

[Euro21a] European Commission (2021, March 17). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Retrieved from: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

[Euro21b] European Commission (2021). The Digital Services Act Package. Retrieved 2021, May 10, from: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

[Feis21] Feis, A. (2021, April 11). Google's 'Project Bernanke' gave titan unfair ad-buying edge, lawsuit claims. *New York Post*. Retrieved from: https://nypost.com/2021/04/11/googles-project-bernanke-gave-titan-unfair-ad-buying-edge-lawsuit/

[Geig21] Geiger, G. (2021, January 5). Court Rules Deliveroo Used 'Discriminatory' Algorithm. Vice. Retrieved from: https://www.vice.com/en/article/7k9e4e/court-rules-deliveroo-used-discriminatory-algorithm

[Lipt18] Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31-57.

[Quac20] Quach, K. (2020, July 1). MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs. *The Register*. Retrieved from: https://www.theregister.com/2020/07/01/mit_dataset_removed/

[Rijk19] Rijksoverheid (2019, October 8). Richtlijnen voor het toepassen van algoritmes door overheden. Retrieved from: https://www.rijksoverheid.nl/documenten/rapporten/2019/10/08/tk-bijlage-over-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid

[Verm18] Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7). IEEE.

About the author

Dr. Alexander Boer is a senior manager at KPMG Trusted Analytics. He is an expert in Artificial Intelligence, law, and risk management, and holds a PhD degree in Artificial Intelligence and law from the University of Amsterdam. At that university he worked as an Artificial Intelligence researcher for two decades, applying Artificial Intelligence technologies to practical and theoretical problems in the field of law.