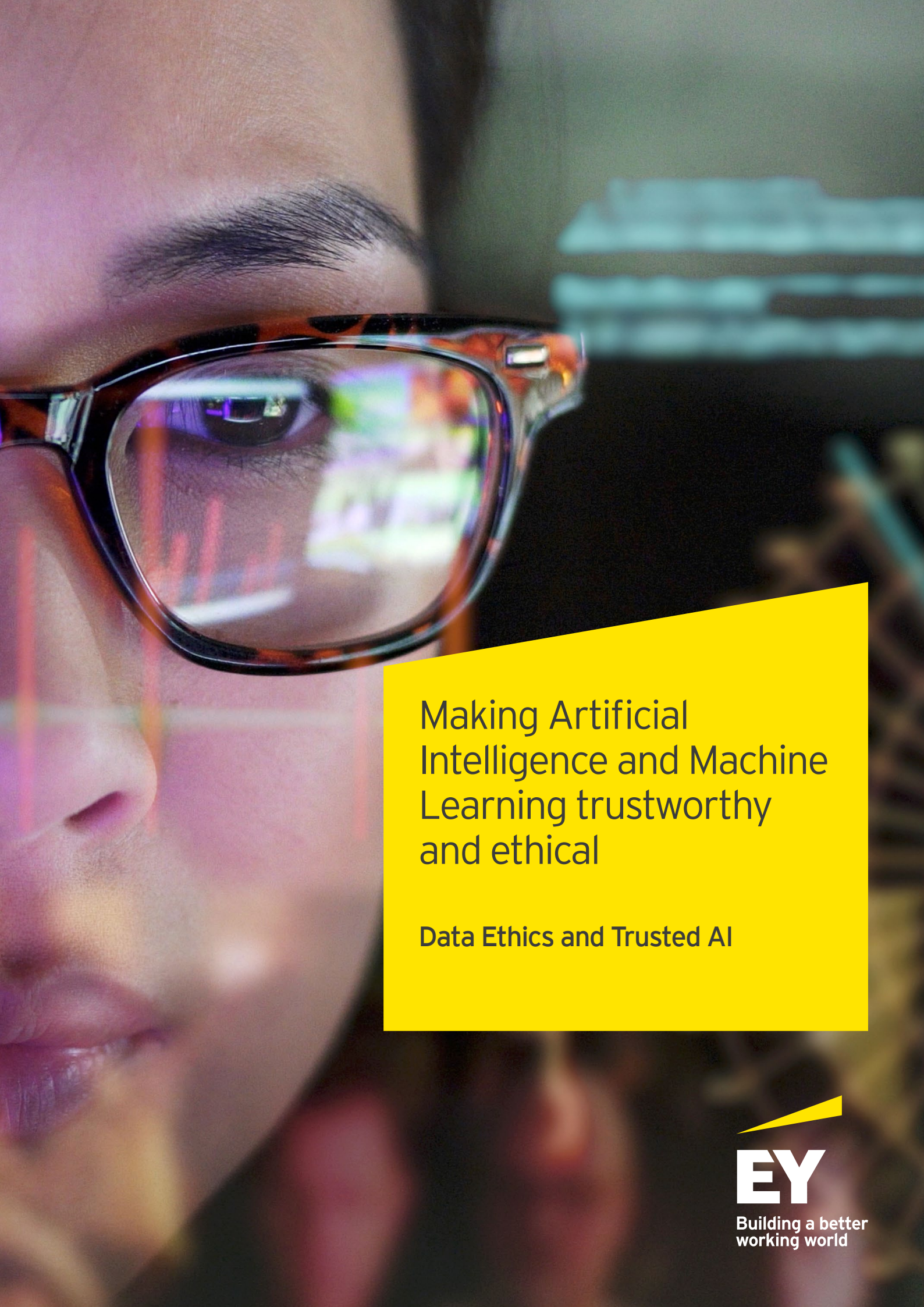


Making Artificial Intelligence and Machine Learning trustworthy and ethical

Data Ethics and Trusted AI

EY
Building a better working world

Executive summary

With artificial intelligence (AI) and machine learning (ML) comes the promise of huge advances for companies, society and even humanity as a whole. However, to reap these benefits, we face a double challenge. First, we need to make it all work on a technical level beyond the toy models and, second, we have to avoid major incidents with its use that could lead to widespread rejection of this new technology.

This paper focuses on the second question: **how to make AI and ML trustworthy and ethical?** While there may be many unknowns on how to realize this, the good news is that it is a process for which the relevant steps can be laid out.

As with any technology, it is the use of AI and ML that defines the risks. While the complexity of the AI approach clearly has an impact on the robustness of the application, the goal it is set to achieve is also a key aspect to take into account. For example, the same image recognition algorithm used to detect faulty widgets in an industrial production process is significantly less risky than when used to detect situational stress of a job applicant performing an online assessment. Next to that, the full value chain, from data collection to usage must be considered.

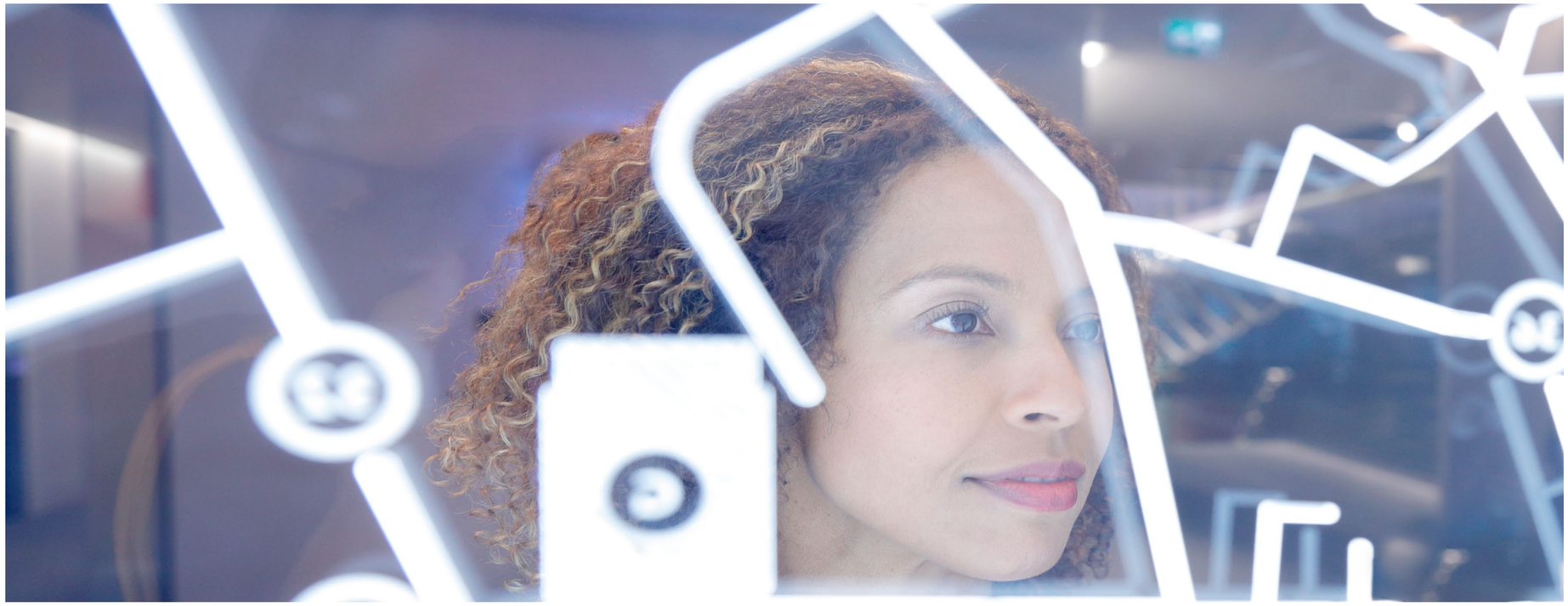
Fortunately, **a differentiated approach is possible, where the most onerous quality, control and governance measures are focused on the use cases that represent the highest risk.** In this regard, we may need multiple risk lenses, with for instance the EU AI Act ("AIA") prioritizing harm to humans in its risk classification, while companies may also consider other applications to be of highest risk on the basis of financial or reputational loss potential. Despite the extra complexity it brings, this risk-based approach is undoubtedly helpful by allowing to strike a balance between going the distance to build trust in the AI applications, and still being nimble enough to innovate. After all, rigorous clinical trials procedures have not stopped the development of new drugs.

Moreover, we know what good looks like in terms of the framework needed to build trust in AI, while the draft AIA provides a way to determine a minimum of no-regret moves in terms of the infrastructure companies need. So there is no reason to wait for the final act in order to proceed: clients, employees, patients, etc. deserve it. Of course, execution details will be important, and will require a good dose of industry-specific considerations.

We're only at the beginning of a multi-year journey to a society where AI applications truly make our life easier, and where we can be relaxed enough that no dangers lurk behind the smile of the friendly avatar or robot.

1

Introduction



The AI revolution creates new risks …

Technological progress has led to an explosion of activity and breakthroughs in the world of AI. The scope of tasks that can be performed by an automated system has expanded, and now includes such things as image recognition and labelling, natural language processing and machine translation, image creation, text writing, continuous signal monitoring and alert generation, etc. In short, a lot of tasks linked to human cognitive capabilities have become technologically feasible in real time. **Efforts are underway to move AI from simple observation and analysis to acting in the virtual world** (all sorts of recommender systems) **and in the real one** (self-driving cars).

The current catch-all name AI captures a bewildering variety of applications. Still, for the most part, AI applications are "merely" new tools in support of existing tasks and duties. A first step is thus to "simply" **adapt existing legal and risk control frameworks to incorporate how this new technology impacts the current situation. However, in many cases, AI also creates totally new risks.** For example, recommender systems and gamification of human-technology interaction actually induce behavior change in human beings, and can create sources of stress unknown before, threatening human autonomy. Next to that, there is the challenge of operational robustness of AI applications deployed at scale. This includes verifying that the AI functions under almost all possible imaginable situations, and that it is safe from cyber-attack.

> "Societal impact and concerns can have significant influence on the speed of AI adoption. They include the way AI will influence humans and societal dynamics, and the need to manage a workforce transition as AI solutions are being adopted.

… in an economic and societal context

A whole ecosystem of companies is trying to capitalize on this intellectual revolution. This ranges from global technology companies to a universe of small startups that use the – often open-source – technology building blocks to build focused, customized applications. Both are necessary to exploit the full potential of AI for the global community. Undoubtedly, the smaller, local companies are important conduits for embedding the relevant AI knowledge into the local economy and allowing for that same local industrial complex to reap its benefits.

Societal impact and concerns can have significant influence on the speed of AI adoption. They include the way AI will influence humans and societal dynamics, and the need to manage a workforce transition as AI solutions are being adopted.

At the human level, **being subjected to automated decision-making is an uncomfortable experience, with or without bias in the algorithms.** AI allows for more continuous monitoring, surveillance, either visually through automatically monitored cameras, or invisibly, using all sorts of data exhaust. GDPR has started to address some of these concerns, and the draft AIA continues to set a clear direction. Beyond observation, the AI-driven insights are increasingly used to nudge or explicitly steer people, their opinions and actions alike. This goes from recommender systems on web shops, over news feed selection algorithms on social media, to fatigue monitoring systems for truck drivers, to real-time instructions (routing) of human operators in logistics. Implicitly or more explicitly, these reduce the human autonomy, and likely increase discomfort. **When becoming exploitative, the AIA labels such uses as "forbidden".**

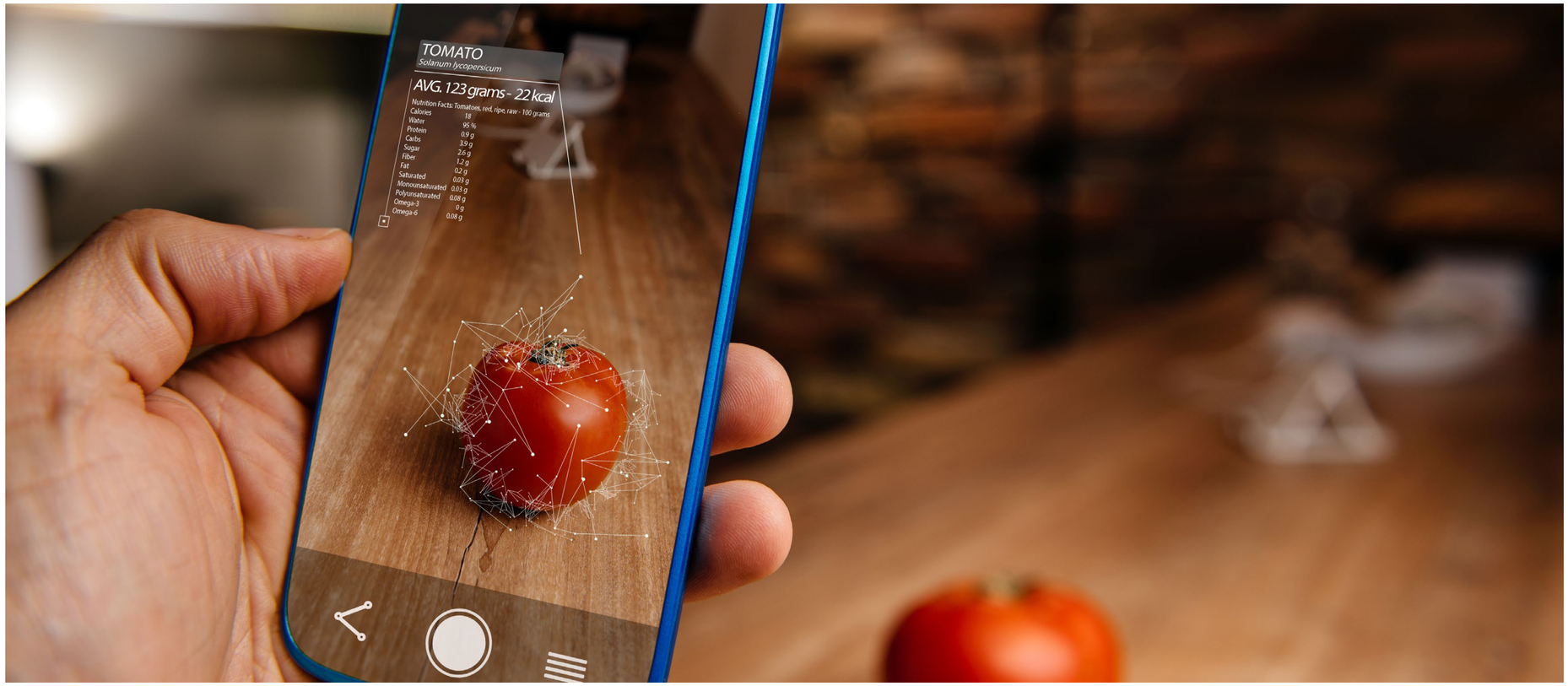
Stakes for companies

In this context, companies have a role to play in helping society realize the promise of AI. Their use of AI should protect individual rights and maximize quality of goods and services.

It is the task of legislators and supervisors alike to create the guardrails for this, all the while creating a stimulating entrepreneurial environment such that investors and entrepreneurs can get adequate return on investment in this new technology.

2

Scope of application



> Two types of steps are necessary to prepare an AI or ML algorithm : collection and curation of data and extraction of patterns.

Before an AI or ML algorithm can be put to use, it must be prepared, "trained", for the task at hand. Roughly speaking, the necessary steps fall into two categories:

- Collection and curation of all sorts of data which reflect (historical) reality
- Extraction (explicitly or implicitly) of some patterns, relationships, knowledge about this reality using one of the different approaches to AI, possibly enriched by a priori business domain expertise of the human modeler

These relationships are presumed to reflect reality, so that when applied to new data instances, they will predict or interpret something about this new reality. In short, **we can see an AI or ML algorithm as an application that uses input data, to provide an output that goes beyond mechanical computation, and that would emulate aspects of human cognition** (perception, judgement, ...).

Let's make this a bit more tangible, using two business applications of a different nature:

- Application A uses automated visual checks to identify decaying fruit on a conveyor belt in a food processing plant (Fruit AI).
- Application B assesses whether a claim in the context of car insurance, is genuine or potentially fraudulent (Fraud AI).

Fruit AI requires data sets of fruit, both fresh and in different stages of being less fresh, under relevant contextual conditions (environmental light, background, ...). Moreover, the different images should be labelled as fresh/not fresh to enable the training. In terms of algorithms, this requires the deep learning convolutional neural networks that have given the industry such a boost the last decade. Relatively little business knowledge needs to be embedded, once a well-labeled data set is available.

Fraud AI builds on a data set of past claims data, with their typical information content such as parties involved, amount of the claim, subject of the claim, etc. Again, the historical data needs to be labelled as (thought to be) fraudulent or not. Models used to estimate a probability that a particular claim is fraudulent can be relatively "simple", and it would not be uncommon to embed business knowledge into their construction to identify well-known fraud patterns.

3

Risks in general – It's the use case that matters

> "As is the case with almost all technologies, it's not the technology itself that is perceived to be risky, it's its use.

Bottom line, it all comes down to one thing: doing the right AI in the right way. Roughly speaking, there are two broad categories of risks associated with AI and ML applications in this regard:

- **Their technical robustness:** does the algorithm always do what it is supposed to do (doing AI right)?
- **Ethical/societal robustness:** doesn't the AI/ML application cause harm to individuals and society? This is sometimes called "the alignment problem" – doing the right AI.

Notice that **these two risk categories are of a totally different nature,** and require different (human) skill sets to address them in the context of a company. **However, many regulations, including the draft AIA, address them together.** Importantly, the different components of the governance and control framework that we describe below allow to mitigate both risk categories at the same time.

Technical robustness covers such things as:

- The quality and relevance of the data set used for training the model
- The stability of the model predictions in function of a changing data set
- The choice of algorithm to perform the analysis
- The accuracy by which it has been implemented in the IT systems
- The overall model performance for instance, false positives/ negatives.

Technical robustness also includes such areas as safety from cyber-attack (e.g., data poisoning), and unintended consequences when used beyond the initial field of application (accidentally or intentionally). Stopping algorithms from jeopardizing individual privacy, either in their inputs or their outputs, is probably the example of ethical robustness that first comes to mind. In the European Union, separate legislation – General Data Protection Regulation (GDPR) – already largely addresses this concern.

However, **the ethical questions** go much wider. They include for instance:

- Does the algorithm (e.g., recommending people for recruitment or release on parole) treat different segments of the population fairly and equitably (absence of bias)?
- Does the algorithm, through its interaction with human beings, imperceptibly alter their behavior, reduce their agency? For example, a recommender system that is mildly risky when it brings books or movies to your attention, reinforcing your existing tastes, can have hugely negative consequences on a liquor, betting or stock trading website, leading to full-blown addiction.
- Does the AI/ML application create psychological pressure, for example because it enables continuous, real-time monitoring of human individuals or because it pilots them around as order picker in a warehouse with tight performance targets?

As is the case with almost all technologies, it's not the technology itself that is perceived to be risky, it's its use. This has consequences on the way the risk framework is set up, it's use case-based.

Let's take a simple example. Imagine a stock trading website that collects browsing behavior to determine its customers' investment risk profile in order to :

- (i) Enable its internal budgeting
- (ii) Tell third-party fund vendors what investor profiles are in its ecosystem
- (iii) Bring relevant investment opportunities to people's attention

All this would fall under GDPR, with consent or legitimate interest appropriately documented, separately for each use case. If the same profile information should subsequently be used for (iv) setting credit lines for the investors, this becomes a high-risk application under the draft AIA, with an altogether more stringent requirement for the internal control environment.

4

There's a minimum of statistics involved

> Increasingly, algorithms are being used to support (profiling) decisions that impact people's lives, such as hiring decisions, access to credit, available educational options, verdicts in the judicial system etc.
>
> The draft AIA labeled such applications as High Risk, subject to the most demanding requirements in terms of governance and internal control.

Beyond the general risks outlined above, it's worth zooming in on two aspects of the risks associated with algorithms. They are a bit more technical (statistical) in nature, but they must be understood by process owners/business owners as they hold the responsibility to make decisions with respect to these points.

The first is deciding the trade-off between False Positives and False Negatives (or other combinations of the classification performance of the algorithm), often named Type I and Type II errors.

If the Fruit AI rejects too much fresh produce (False Positive), the profitability goes down, as perfectly good products go to waste. On the other hand, letting too many fruits, about to turn bad, go through (False Negative) and end up in the hands of the consumer, is detrimental for the company's reputation in the long run. Getting this trade-off right is the most important selection criterion in this case. The fact that the consequential costs are asymmetric and not equally transparent complicates the task. Indeed, while bringing fruit to the dump is an immediate, quantifiable cost, the loss of brand power due to long-term quality issues, is less directly visible or quantified.

Next to that, the topic of algorithmic bias and fairness of outcomes has come to the fore dramatically over the last five years. Increasingly, algorithms are being used to support (profiling) decisions that impact people's lives, such as hiring decisions, access to credit, available educational options, verdicts in the judicial system etc. The draft AIA labeled such applications as High Risk, subject to the most demanding requirements in terms of governance and internal control. In this context, **the outcome of the process supported by the algorithms should be "fair and unbiased", with respect to relevant dimensions**, such as gender, religion, age, etc. What these "relevant dimensions" are, can depend on legal requirements, societal norms or fundamental ethical values.

There are two reasons why this is a very challenging new domain for users of machine learning and AI applications. First of all, different perspectives on what is considered fair can be incompatible when confronted with reality. Consider for example the world of (consumer) credit. One possible way to define fair outcomes in this context would be that men and women applying for credit should have the same success rate. On the other hand, one might argue that applicants with the same salary should have the same likelihood of a positive outcome, as the salary determines the payback capacity.

However, as long as men and women don't have the same salaries, these two objectives of fairness can't be reconciled. Clearly, **which of the definitions of fairness is important for the company is a business choice**.

It's worth keeping in mind that these biases are only reinforced or made more visible by the algorithms. They existed before, and recruiters, judges, credit officers, etc, may have made assessments in the past that were equally biased or unfair. One of the consequences of this is that, **in training AI models, one should carefully assess whether the historical data contain such biases that would subsequently be amplified by the algorithm**. Legally, a principle that is often applied is that individuals should not be subject to automated decisions on the basis of such algorithms without human expert validation. That is certainly a good safeguard. On the other hand, it would be good practice to also have the human experts' decisions challenged by the algorithms to potentially identify their implicit human biases.

From the bias and fairness examples, it's clear that any approach to remedying the situation needs to be holistic in nature, covering the end-to-end process of algorithm development and deployment within a well-defined business context. Let us now describe the robust set-up that is necessary to achieve this.

5

Components of the governance and control framework

In different organizations, there is currently a bit of a debate to determine whether AI/ML is a separate, new risk category, or whether it should be treated within the context of the risk profile of the business processes that rely on them. Considering that it's the usage that matters, there is a case to be made for the latter approach. However, the inherent technicality of the topic may necessitate to treat AI/ML as a separate risk category in order to make progress in the current context.

Whatever one's point of view in the above debate, it's clear that AI changes the organization's risk profile, sometimes significantly. The good news is that the conceptual framework applied to risk management and governance for different risk categories, easily applies here too.

> "In essence, "explainability" makes the model outcome reasonable to human users by clarifying the causes.

- First of all, **does the organization have a complete overview of where AI applications are being used?** Does it have an inventory of where things could go wrong with AI/ML? Some uses are very visible (e.g., in a recruitment process), some may be more hidden (e.g., Natural Language Processing enablers in all sorts of intelligent automation). Does the inventory contain adequate information on the intended use of the AI so that its actual performance can be assessed against this objective?

- Given the understanding of the intended use, and the approach and data set, **does the organization have a clear way of identifying the use cases that present the highest risk**, in order to focus the control efforts (development testing, curation of training data sets, understanding of the algorithm, documentation, …) on these applications? In other words, is there a well-defined approach to risk assessment of the different use cases? It's important to emphasize that the biggest efforts are only required for the highest risk cases, which are likely to be only a (material) subset of the entire inventory, probably between one third and half of the use cases at worst.

- **Are the necessary risk-mitigating actions identified and implemented?** Often, a separate algorithm validation exercise, in particular for the high-risk ones, helps to identify the weakest links, or assess the adequacy of the mitigating actions (the Zoom 2 section lays out what such an algorithm validation could entail). These controls can be one-off or of low periodicity (e.g., data set vetting, post model calibration) or can consist in ongoing performance monitoring.

- Finally, **does the organization have adequate internal reporting**, such that the stakeholders at different levels have a good insight into the company's AI risk profile, and into the effectiveness of the control framework.

In this context, we need to pay special attention to the "explainability" of model outcomes, in particular for the more complex algorithms in ML and AI, such as deep learning or ensemble approaches. The two examples used previously, illustrate that this ability to explain is only relevant in certain use cases.

Being able to justify in real time why the machine thinks a certain piece of fruit is rotten, is of limited added value, once the system is put in production. It's only useful on a periodic basis, when the performance is evaluated, as it can help remediate findings.

However, for the Fraud AI, which has potential legal or customer experience consequences, it is important to be able to explain the model's decision procedures to third parties (such as the customer), effectively in real time.

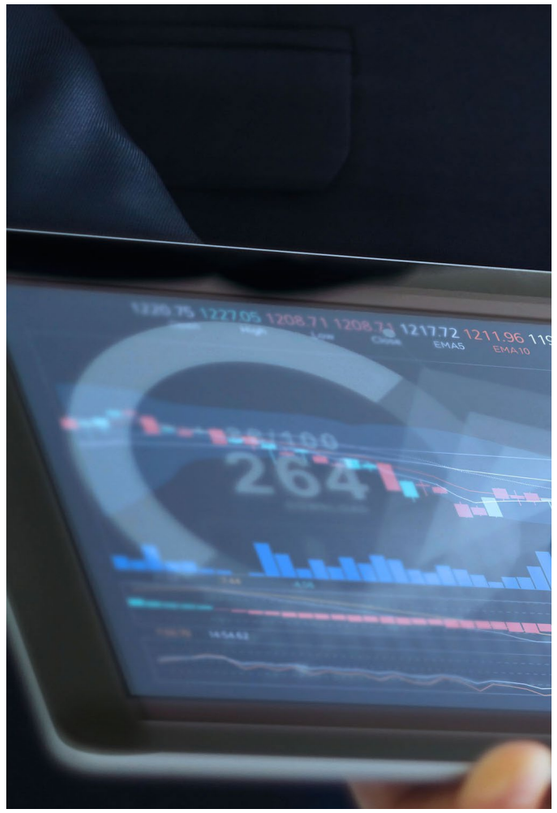
"Explainability" algorithms are themselves models that aim to provide insight into how the more opaque algorithms arrived at their "conclusion". Typical ingredients are:

- Rank orderings of the data fields that were important in the analysis, or
- Surrogate linear models that can be used to replicate the more complex models for specific decisions, and that are readily interpreted.

In essence, "explainability" makes the model outcome reasonable to human users by clarifying the causes. Given that such tools are algorithms in their own right, the development and documentation principles laid out above, equally apply. One should also embed these "explanation tools" into the process flow, precisely because they drive the human-algorithm interaction. When this needs to be done in real time (like, for instance, in predictive policing), the technology demands are high. We will address this topic in the Zoom 1 section on ML Ops.

6

Conclusion



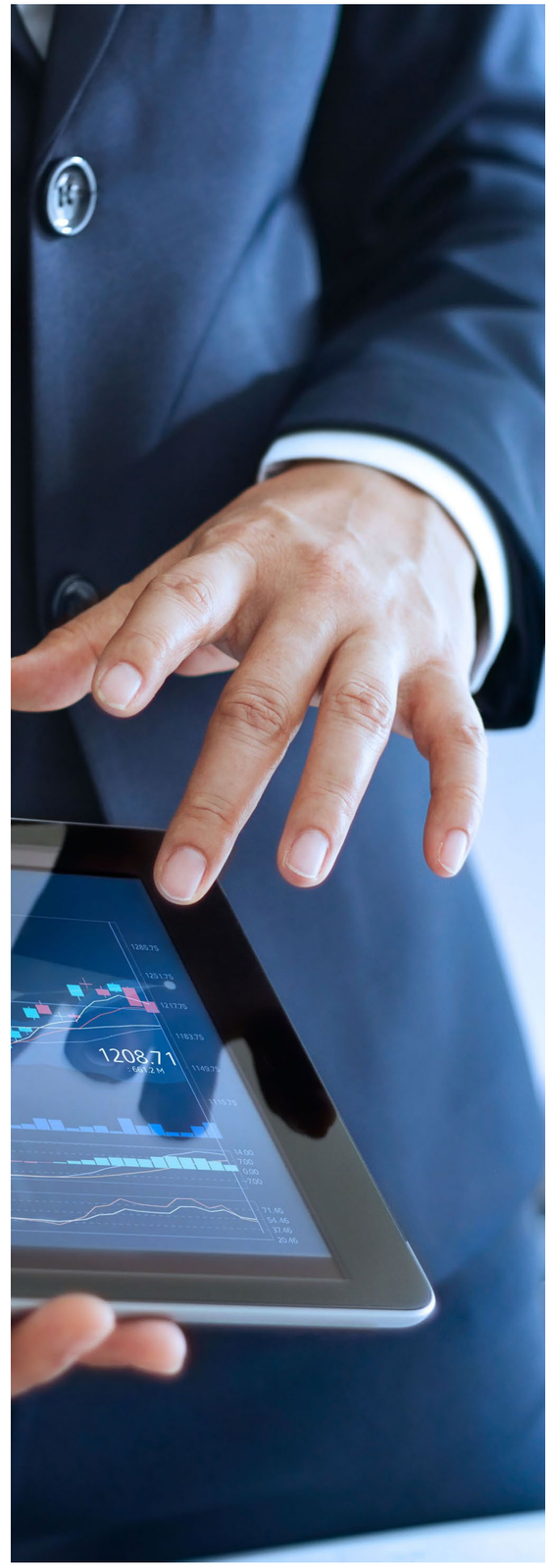
Trust in the use of Big Data, advanced analytics and AI will depend on the way it's put into operation. It is quite clear what general design principles we need in order to put in place a framework that minimizes the frequency and the impact of AI-related problems. These principles are also increasingly codified in legislation such as the draft AIA.

However, implementing them requires paying attention to the right details at all levels, from business specification, over algorithm design, to making it operational in a robust Information Communication Technology (ICT) infrastructure. Getting this right is a collective learning journey for everybody involved in this new and evolving technology.

Now is the time to start taking a series of no-regret moves such as determining the AI use cases that can put the organization at risk, and to start learning how to control and mitigate these risks. This will also be an important first step paving the way for transparent and credible communication with all stakeholders around the topic. Doing the right thing, and explaining what was done and why, are prerequisites for building sustainably the appropriate level of trust.

7

Zooms



Zoom 1: ML Ops

“

Models in production often underperform relative to the prototype or even outright misbehave.

Zoom 2: Detailed algorithm audits or validation

“

Tried and tested approaches exist, to assess the functional robustness and performance of AI and ML models.

Zoom 1
ML Ops

It's not enough to develop AI/ML models, they have to be put in production, often in very demanding circumstances where high volumes of data need to be handled in near real time, all the while verifying that the model keeps on doing what is was trained to do. **Models in production often underperform relative to the prototype or even outright misbehave due to different reasons**, including:

- **Data drift**: the statistical properties of the data, inference, concept or label change over time, which results in your model losing its relevance to the current situation.
- **Biases**, that were within acceptable limits in the prototype, can be more dangerous due to ongoing re-training or because of different populations to which the algorithm is applied.
- **Security breaches** that are exploited to change the model behavior. This can include adversarial attacks where a constantly-updating learner is poisoned with erroneous data.

Should any mishap occur, it will be important to have an **audit trail**, allowing the analysis of what went wrong and take measures to avoid repetition.
Monitoring systems typically cover three different layers of performance, namely:

- business performance
- infrastructure and application performance
- the actual ML model performance

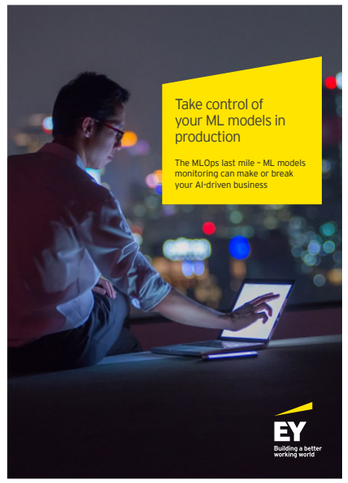
Business monitoring offers stakeholders (product managers, business analysts,...) a holistic overview on their business metrics, in order to identify unusual variances and get alerts to take action. **Infrastructure and application monitoring** is the traditional software/infrastructure monitoring tackling well-known issues such as ML model request latency, access to memory and CPU, etc. Resiliency and scalability are major considerations in this regard. Finally, **ML model performance** provides a solution to monitor the ML models running in production, allowing you to address and control the issues mentioned above.

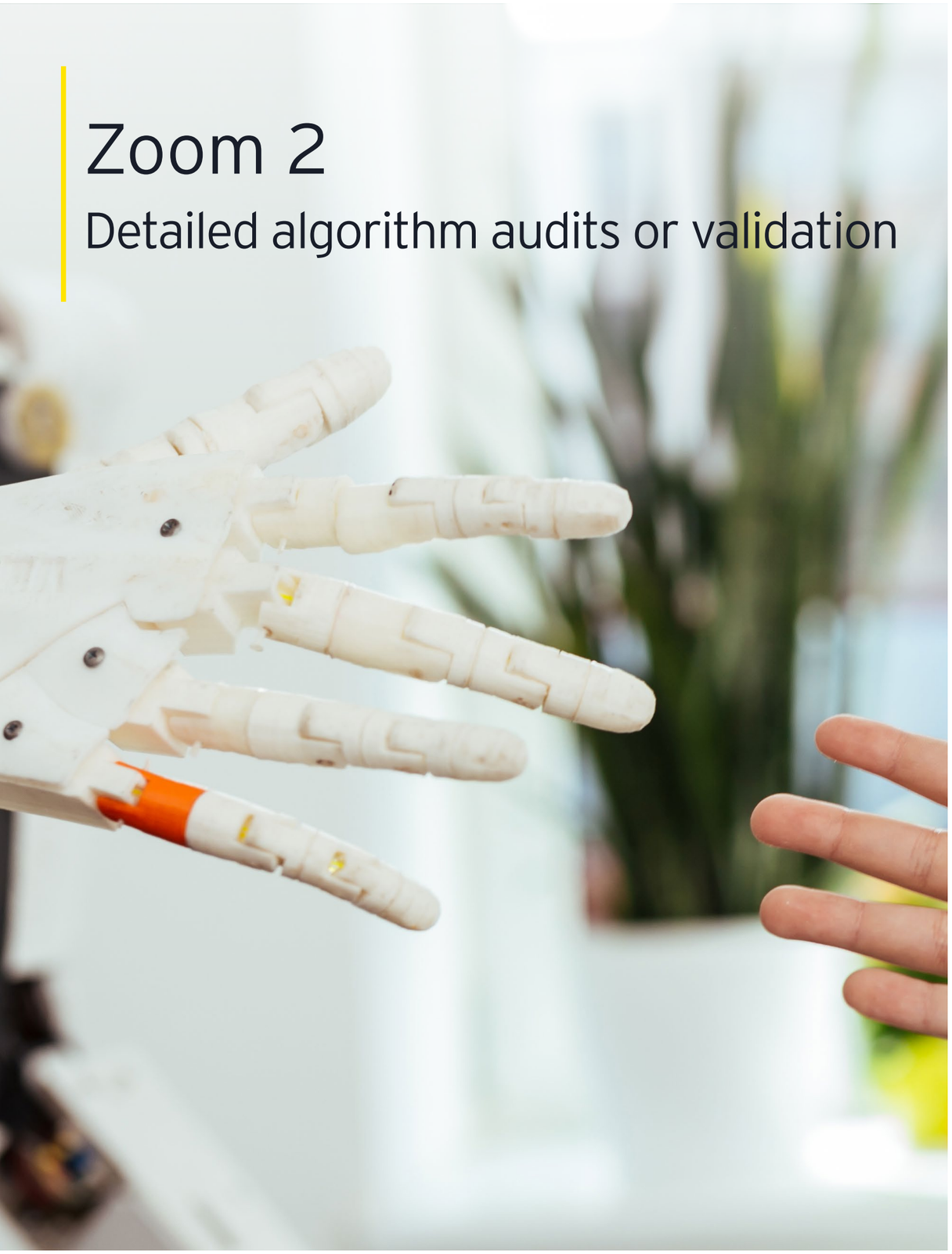
Having defined the required monitoring capabilities, the challenge remains on **how to operationalize** it. A variety of approaches are currently observed in the market. They can be classified along two axes:

- Is the monitoring capability integrated in a ML platform or not?
- What is the scope of the monitoring capability?

In summary, it should be clear then that **setting up the appropriate monitoring framework is in itself an important project**: what are the quantities that should be monitored, and how will this be done technically?

For an overview of ML monitoring space - clusters, see the MLOps paper "Take control of your ML models in production" by Patrice Latinne and Amir Krifa (available upon request).





Zoom 2

Detailed algorithm audits or validation

Several industries, including pharma, banking and insurance, already have **quality and control standards when it comes to the use of mathematical models and statistical inference techniques** that support decisions that are important for the company and for society. Tried-and-tested approaches therefore exist, to assess the functional robustness and performance of AI and ML models.

For instance, since the introduction of Basel II in the banking industry in the early 2000s, **banks have set up independent model validation units**, that follow increasingly well-defined workplans to be able to sign off on the reliability of credit capital models. The templates developed in this context can serve as basis for AI assessments, although, of course, the specific model risk profile of the AI and ML models must be taken into account.

The typical areas addressed by a model audit or algorithm validation are:

Validation considerations from methodological point of view

Data

Sampling and population exclusion → Data partitioning → Data pre-processing → Training data representative-ness → Feature exclusion → Data quality → Data bias

Conceptual soundness

Feature engineering and selection → Hyperparameter specification → Alignment and coverage of assumption → Overfitting and underfitting → Libraries and packages selection

Performance assessment

Performance tests → Benchmarking → Model stability → Data and concept drift → Dynamic models → Ongoing monitoring

Model technical limitation

Data → Skill shortage → Cost → Continuous supervision → Bias

Family-specific components

Trad/supervised learning

- Performance measurement
- Imbalanced data
- Different types of Deep NN and their architectures
- Semi-supervised learning
- Active learning

Unsupervised learning

- Clustering approach
- Outlier detection
- Performance measurement

Natural language & voice

- NLP system validation
- Performance assessment
- NLP applications validation
- End-to-end adversarial text-to-speech

Reinforcement learning

- Modeling framework
- Exploitation vs exploration
- Hybrid RL methods
- Model stability and convergence
- Overfitting and underfitting
- Offline/online model performance measurement

Your EY contacts

Patrice Latinne
Partner - Data & Analytics
Ernst & Young Consulting
patrice.latinne@be.ey.com
+32 472 980 719

Frank De Jonghe
EY EMEIA Trusted AI leader
frank.De.Jonghe1@uk.ey.com
+32 479 90 67 59

Amir Krifa
Executive Director - Data Science & AI
Ernst & Young Consulting
amir.krifa@be.ey.com
+32 478 992 478

EY | Building a better working world

EY exists to build a better working world, helping create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

EYG no. 010804-21Gbl