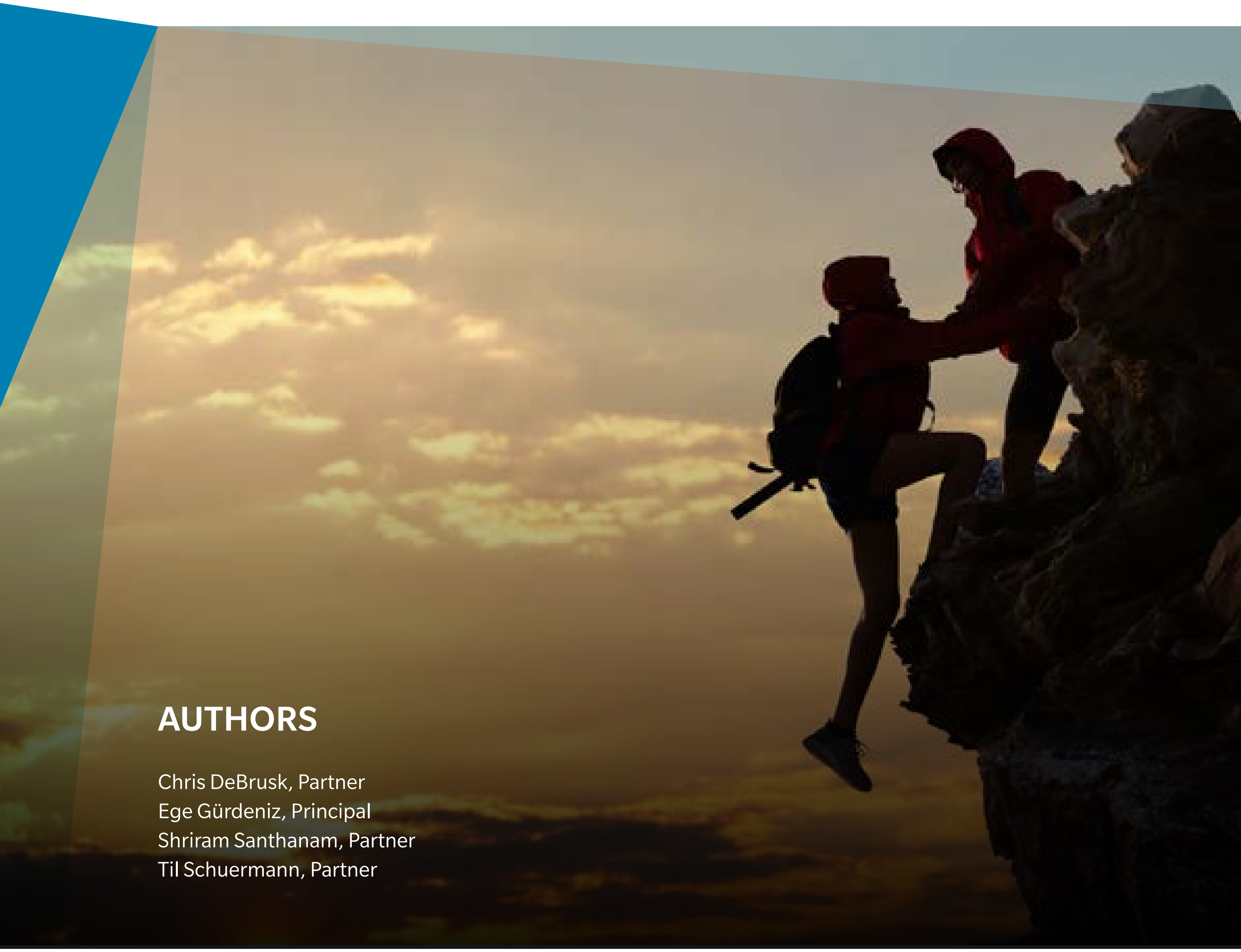
OLIVER WYMAN

TRUSTING THE MIND OF A MACHINE



AUTHORS

Chris DeBrusk, Partner
Ege Gürdeniz, Principal
Shriram Santhanam, Partner
Til Schuermann, Partner

MARSH GUY CARPENTER MERCER OLIVER WYMAN

MARSH & McLENNAN
COMPANIES

INTRODUCTION

"If you can't explain it simply,
you don't understand it well enough"

-Albert Einstein

As the adoption of applications that leverage complex machine learning grows, so do concerns around humans' ability to sufficiently understand and explain the decisions made and actions taken by machines. This concern has been particularly pronounced in areas where the lack of understanding around modeled output can have a real, negative impact on customers such as unfair treatment of loan applicants within financial services or misdiagnosis of patients within health care.

Various terms such as Artificial Intelligence (AI) "explainability," "transparency" and "interpretability" have been used by different groups and organizations to articulate this challenge. However, the fundamental issue boils down to our ability to trust the output produced by the machine: to make a significant decision that impacts others based on a piece of output, we must sufficiently trust the output, and to sufficiently trust the output, we must:

1. Know that the output is accurate
2. Sufficiently understand how and why the output was produced

Most institutions have independent review frameworks and qualified testers that make sure the output produced by a machine is accurate and appropriate (for example, model risk management functions at financial institutions). Knowing that a qualified third party has reviewed and certified a machine for use does establish some level of trust in the system. However, independent review does not necessarily help others understand the machine: a high-performing machine that has been independently reviewed and certified by a highly qualified team of computer scientists can still be a complete mystery to parties that are impacted by the machine such as users and customers.

Indeed, not every machine needs to be understood and explained (and trusted) at the same level. Depending on the use and purpose of the machine, who is impacted and how, and other factors such as regulatory requirements, each machine can have a different requirement for the level of trust. As a result, an effective solution to the trust issue must account for these different factors. In this article, we present a framework that institutions can use to establish trust in the "Machine-Human Ecosystem", and enable the responsible and large-scale adoption of machine learning applications.

Artificial Intelligence or AI, is a branch of Computer Science. **Machine Learning**, is a field of study within AI that focuses on a particular class of algorithms. Arthur Samuel, who coined the term "machine learning" in 1959 defines it as **"the field of study that gives computers the ability to learn without being explicitly programmed"**.

With the rollout of the General Data Protection Regulation (GDPR) by the EU and overall heightening supervisory and public expectations around data privacy and traceability, being able to explain a machine's output to customers and clients has become a legal obligation (at least for firms doing business in the EU). We recommend companies take steps now to establish a framework that will allow them to meet these increasing requirements and expectations in a timely manner.

THE CHALLENGE

Machine learning algorithms can take many shapes and forms, and vary in complexity. As a result, our ability to understand the output produced by a machine can be impacted in different ways depending on the specifics of the learning algorithm. However, at a high-level and in simplified terms, understanding the output of machine learning applications can be challenging due to three key factors:

- **Self-programming:** Learning algorithms program themselves, adjust their parameters, and define their own rules by mapping inputs to outputs following certain mathematical processes and objective functions as opposed to following explicit rules and logic programmed by humans
- **Input-to-output complexity:** The large number of model components, layers and parameters which are adjusted by the machine as part of the learning process, and the complexity of interactions between these self-adjusting parameters makes it extremely challenging to attribute and link inputs to outputs
- **Network effect of machines:** The output of a self-programming machine can be used as the input to another self-programming machine without any human involvement, creating a complex and opaque network of machines that is difficult to comprehend

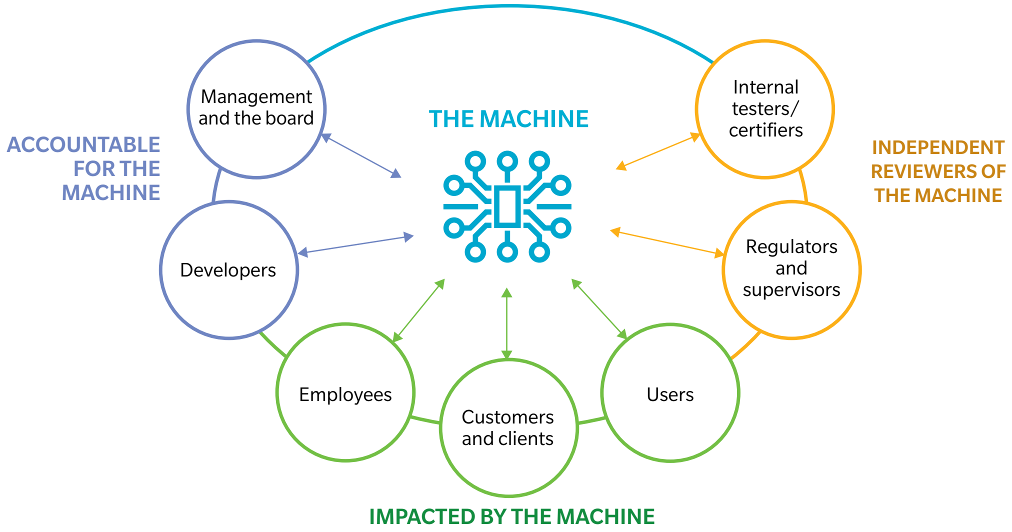
Model explainability is a continuum which is correlated with model complexity. For example, a simple regression is much easier to understand and explain than a multi-layer neural network. However, while simpler models in general are easier to explain, they also in general have lower performance (for example, lower accuracy of predictions). Therefore, as companies attempt to solve increasingly complex problems with increasing accuracy, they will need to use increasingly complex approaches such as deep neural networks that can have tens of hidden layers and thousands or millions of parameters with non-linear interactions, which humans cannot intuitively or immediately understand. With increasing complexity, trusting the machines will become increasingly difficult.

Google's Deep Mind neural network reportedly has around **10 billion neurons**. The largest artificial neural network, developed by Digital Reasoning, reportedly has around **160 billion neurons** – or parameters. In comparison, the human brain has in the order of **80-100 billion neurons.**

WHO NEEDS TO UNDERSTAND?

The "Machine-Human Ecosystem" comprises various groups of people with different levels and types of interactions with the machine. As a result, each group may have a different level of need and ability to understand the output produced by the machine. The question then becomes "how much does each group need to understand?"

Exhibit 1: The Machine-Human Ecosystem


ACCOUNTABLE FOR THE MACHINE
Management and the board
Developers
THE MACHINE
Internal testers/ certifiers
Regulators and supervisors
INDEPENDENT REVIEWERS OF THE MACHINE
Employees
Customers and clients
Users
IMPACTED BY THE MACHINE

It is not reasonable or possible to expect that every group will have a complete understanding of all the mechanics and mathematics behind the machine, and can explain its every aspect. On the other hand, we also cannot accept a state where every group has a very limited and superficial understanding of the output. If the number of traditional models used in companies today is any indication, this problem is further complicated by the fact that there soon may be an inventory of hundreds or even thousands of machines making critical decisions in any given company.

Given the differing needs and ability to understand as well as the variety and number of machines that will likely exist, institutions cannot follow a one-size-fits-all approach to establishing trust in the ecosystem. We propose a "tiered trust framework" which is based on the classification of the machines following some objective criteria.

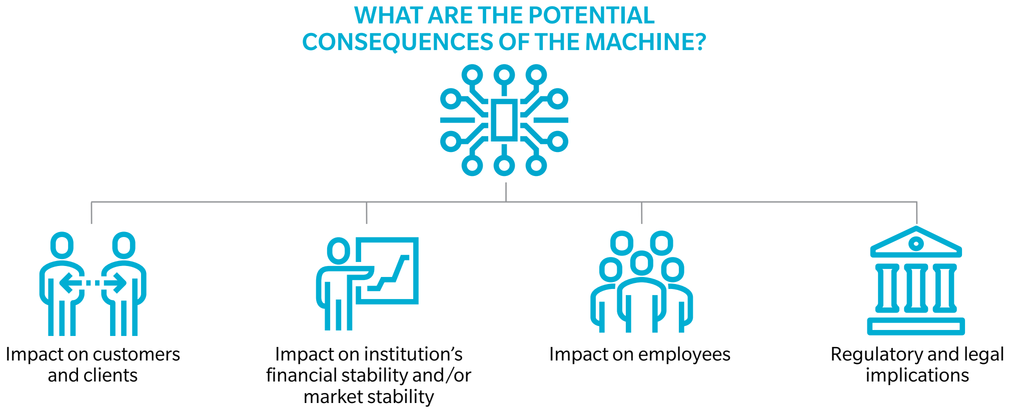
The pre-AI Google Translate used a set of **human-defined rules and logic** to translate texts phrase-by-phrase. The current Google Translate uses **neural machine translation,** which translates whole sentences by learning the semantics of languages as opposed to following set rules. This more complex approach **reduced translation errors by 60%.**

THE TIERED TRUST FRAMEWORK

Not every machine is the same, and therefore not every machine needs to be understood and trusted at the same level. If the machine is consequential in any way, such as impacting clients and customers, fulfilling legal requirements or impacting the institution's financial well-being, it should be placed higher in the tiered trust framework, and the requirements around understanding the model output should be appropriately higher for each impacted group. However, if the machine is inconsequential, for example, used only for a minor and limited internal task to increase a team's efficiency, then it should be placed lower in the framework, with commensurate requirements.

As a starting point, we recommend machines be assessed along four impact dimensions to determine the potential consequences of the machine output.

Exhibit 2: Assessing a machine's impact



WHAT ARE THE POTENTIAL CONSEQUENCES OF THE MACHINE?

Impact on customers and clients

Impact on institution's financial stability and/or market stability

Impact on employees

Regulatory and legal implications

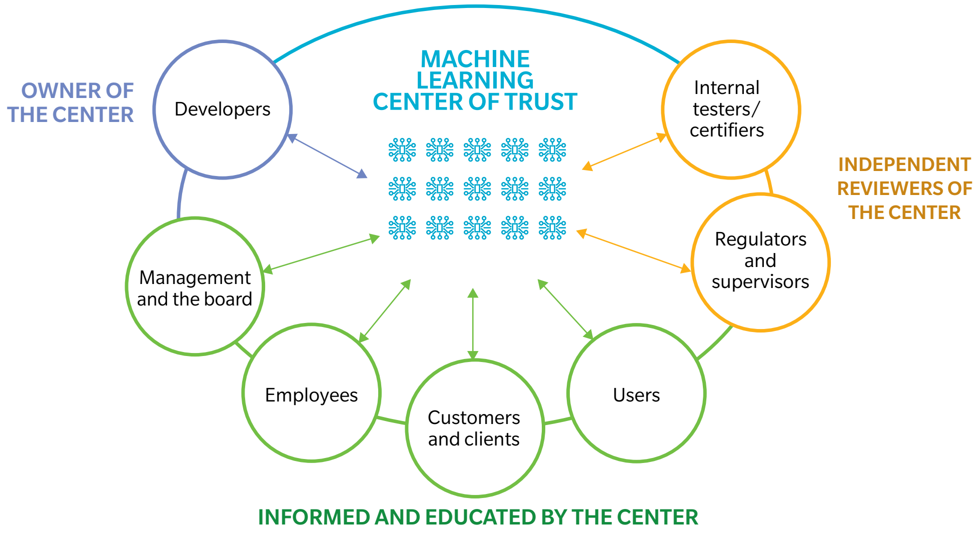
Each institution will have its own thresholds and definitions for the above categories, but ultimately each machine will fall on a spectrum of "high to low" across the various dimensions, requiring a variety of efforts and artifacts to ensure the impacted parties have a sufficient understanding of the machine output. However, who will be responsible for developing and owning this framework, putting the machines through the framework, and ultimately explaining the model output to the impacted parties?

"Clustering" is a common form of machine learning where an algorithm can consume massive datasets (for example, millions of daily customer transactions) and identify hidden patterns and trends **not visible to the human eye.**

THE MACHINE LEARNING CENTER OF TRUST

Given the large number of machines and impacted parties involved as well as the need to follow a consistent methodology, this challenge is best addressed centrally. Thus, our recommendation is to designate the existing machine development function (Head of Data Science, Head of AI or Head of Analytics) as the "Machine Learning Center of Trust", which would be responsible for owning and executing the tiered trust framework and developing the necessary artifacts to help impacted groups understand and trust the machine.

Exhibit 3: The Machine Learning Center of Trust


MACHINE LEARNING CENTER OF TRUST
OWNER OF THE CENTER
Developers
Internal testers/ certifiers
INDEPENDENT REVIEWERS OF THE CENTER
Regulators and supervisors
Management and the board
Employees
Customers and clients
Users
INFORMED AND EDUCATED BY THE CENTER

The Machine Learning Center of Trust ("the Center") would have the following key responsibilities with respect to explaining the machine:

- **Testing:** Running a host of quantitative tests to assess model input significance and impact on output – for example:
  - Creating and reviewing partial dependence plots which indicate how each feature (variable) impacts the model prediction
  - Applying testing techniques that assess the impact of input perturbations on output (for example, LIME[1])
  - Reviewing feature importance scores which indicate how useful or valuable a feature (variable) was in determining the output
  - Reviewing variable coefficients

Machine translation traditionally **relied on English as the bridge** between the translated languages. On the other hand, **Google's Neural Machine Translation** algorithm taught itself to translate without using English as the bridge. The **neural network accomplished this by developing its own "machine language"** that allowed for efficient translation between multiple languages.

[1] "Local Interpretable Model Agnostic Explanation" (LIME) is a technique that fits a linear (and thus easily interpretable) model on local perturbations of the complex model to explain its predictions

- **Data review:** Walking through the data sourcing methodology, and tracing back the data used to train the model in order to identify and remediate any potential areas of bias
- **Documentation:** Creating user friendly documentation that synthesizes the results of quantitative tests and any other qualitative assessments that were made (for example, contrastive explanations of the output), and providing a non-technical and intuitive explanation of the drivers behind the model output
- **Procedures:** Defining and implementing machine development standards and procedures to make sure all machines are developed in a transparent and consistent way, and the output of the machine is replicable by independent third parties
- **Monitoring and reporting:** Monitoring model inputs and outputs on an ongoing or regular basis at the appropriate frequency (depending on the tier of the machine), and reporting the results to the relevant parties (for example, management)
- **Training:** Designing and executing targeted training programs, workshops and communications (internal or external). For example, the roll-out of a high importance machine could be accompanied by an appropriate workshop to educate the relevant parties on the new machine
- **Customer support:** Providing customer/employee support on questions related to machine-related inquiries. For example, a customer asking why their credit application was rejected by a robot, or a sales person asking why the machine is recommending they sell a particular product to a client

It will be up to the Center to determine how the above elements will be applied for each tier. However, the Center's decisions and framework will ultimately be independently reviewed and challenged by a third party to ensure the approach is appropriate such as internal audit, the risk management function and supervisors.

CONCLUSION

The potential benefits of successfully using machine learning at scale are numerous and well covered by industry publications, academic papers and mainstream media alike. New use cases, applications and experiments appear daily, further adding to the excitement and optimism around what machine learning can deliver for companies and consumers.

However, the absence of trust in the Machine-Human Ecosystem will likely inhibit the large-scale adoption of machine learning as the risk of unintended negative consequences will be too great, and organizations may not have the appetite to face the potential regulatory, legal, ethical or financial consequences. To avoid this roadblock on adoption, institutions should start designating their own version of the "Machine Learning of Trust" and begin rolling out the associated guidelines and requirements now.

Deep Blue, the chess-playing computer that beat Gary Kasparov, followed **hard-coded rules written by a human** (relatively easy to observe and understand). **AlphaGo**, which beat Chinese Go Master Ke Jie, self-taught the game by watching hundreds of thousands of games and millions of moves using a deep neural network (very difficult to observe and understand).

Oliver Wyman is a global leader in management consulting that combines deep industry knowledge with specialized expertise in strategy, operations, risk management, and organization transformation.

For more information please contact the marketing department by email at insights.digital@oliverwyman.com or by phone at one of the following locations:

AMERICAS	EMEA	ASIA PACIFIC
+1 212 541 8100	+44 20 7333 8333	+65 6510 9700

AUTHORS

Chris DeBrusk, Partner
Chris.DeBrusk@oliverwyman.com

Ege Gürdeniz, Principal
Ege.Gurdeniz@oliverwyman.com

Shriram Santhanam, Partner
Shriram.Santhanam@oliverwyman.com

Til Schuermann, Partner
Til.Schuermann@oliverwyman.com

The information and opinions in this report were prepared by Oliver Wyman. This report is not investment advice and should not be relied on for such advice or as a substitute for consultation with professional accountants, tax, legal or financial advisors. Oliver Wyman has made every effort to use reliable, up-to-date and comprehensive information and analysis, but all information is provided without warranty of any kind, express or implied. Oliver Wyman disclaims any responsibility to update the information or conclusions in this report. Oliver Wyman accepts no liability for any loss arising from any action taken or refrained from as a result of information contained in this report or any reports or sources of information referred to herein, or for any consequential, special or similar damages even if advised of the possibility of such damages. The report is not an offer to buy or sell securities or a solicitation of an offer to buy or sell securities. This report may not be sold without the written consent of Oliver Wyman.

OLIVER WYMAN