Building trust

What human leadership teaches us about developing trustworthy technology

Sophisticated technologies are impacting more and more aspects of our lives. Artificial intelligence, for example, is increasingly used to make important decisions: whose résumé gets them a job interview; who qualifies for a loan; whether a person caught on camera committing a crime matches a photo from a database. **Technology systems are trusted to make or influence decisions with major implications for individuals and for society.**

Or to be more precise, these systems are trusted by the organizations that make and use them. But are they trusted by the company's individual employees, customers, suppliers, and society at large—and what happens when they aren't?

What does it take to build and maintain trustworthy systems?

Table of contents

For guidance, organizations can turn to the same qualities that inspire trust in human leaders, and see how they map to technology systems. Leadership development consultancy Zenger/Folkman analyzed the 360-degree feedback assessments of 87,000 business leaders.[1] They define three key elements that help build a foundation of trust in individual leaders:

01

Positive relationships: Trustworthy leaders stay abreast of others' concerns and balance their results with others' needs

02

Good judgement and expertise: Trustworthy leaders make wise decisions, and others seek out their opinions and knowledge

03

Consistency: Trustworthy leaders "walk the talk" and keep their promises

We think these areas can also serve as guiderails for building a foundation of trust in technology-driven systems. Organizations can commit to trustworthy systems by developing these characteristics in the technology they use, coupled with strong governance and commitment to ethical oversight.

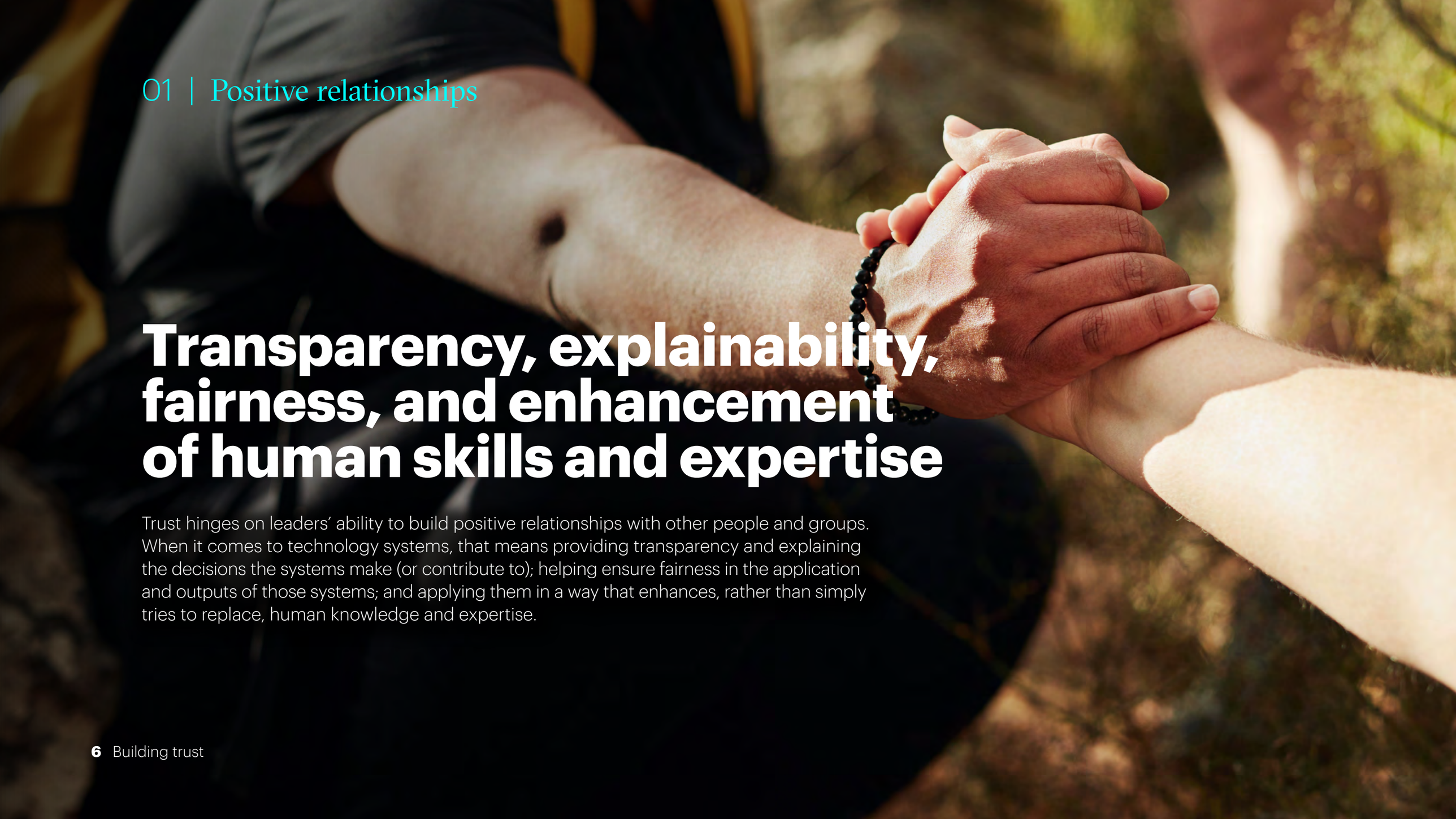
The importance of trustworthy systems

With how pervasive technology has become in business, companies can only be successful if people have confidence in the technology systems they use.

But trust is difficult to build and very easy to lose. One wrong move with an innovative new technology could mean the end of a longstanding customer relationship or community partnership, or increased scrutiny in the form of regulations or legal penalties.

On the other hand, there's a huge opportunity for companies that take the lead in building trustworthy systems. They can achieve greater reach and influence among existing customers and partners as well as society. They can get a head start on meeting regulatory requirements and staying in compliance with shifting legal landscapes. And perhaps most importantly, they can build a foundation of trust, so that people are willing to join them when they innovate in new spaces. Together, these opportunities can help create direct business benefits in terms of brand value, customer retention and business growth.

What does it take to be a tech leader that consistently delivers trustworthy systems? Let's take a closer look at what inspires trust in people. Zenger/Folkman's findings highlighted **positive relationships, good judgement and expertise**, and **consistency**. Read on to see how we think companies can develop and build these qualities into the technology they use, creating trustworthy systems.



01 | Positive relationships

Transparency, explainability, fairness, and enhancement of human skills and expertise

Trust hinges on leaders' ability to build positive relationships with other people and groups. When it comes to technology systems, that means providing transparency and explaining the decisions the systems make (or contribute to); helping ensure fairness in the application and outputs of those systems; and applying them in a way that enhances, rather than simply tries to replace, human knowledge and expertise.

Transparency and explainability

If you can't understand the reasons for someone's decision, it's difficult to trust that what they're saying is correct. This is exactly what can happen when tech-driven systems' inner workings are a mystery—when the systems act as "black boxes" that no one can see inside. It's a particular concern with artificial intelligence, which is increasingly used to drive decisions that affect people's lives. Many AI systems in use today were not designed to provide explanations about why they generated a particular decision or output.

This is also what makes it possible to more easily adjust or change the system if its decisions are biased or incorrect—a key element of responsible use.

Advancing explainable AI

We've helped clients use a type of explainable AI which shows the minimum changes in inputs it would take for a particular model to reach a different outcome: counterfactual explanations. Imagine that someone has been turned down for a loan. The AI would help them understand by how much they would need to change inputs—for example, an increased salary, reduced loan amount, or improved credit rating —to change the decision from a rejection to an approval. This approach can bring explainability even to existing systems that were not designed to generate explanations for their decisions.

Explainable AI can help ensure that systems' decisions are accompanied by clear explanations, making it straightforward for people to understand how and why a system has reached a certain decision.

Fairness

Trustworthy leaders aim for fairness in their decisions, and it shows in their outcomes. Trustworthy systems must do the same.

For organizations, that means quantitatively assessing the fairness of the technology solutions they use, including everything from their data and traditional analytics and algorithms through to more sophisticated artificial intelligence systems. It's important to look at how these systems impact people —both directly and indirectly—and address unintended consequences.

Critically, organizations must also monitor such systems' performance after they're in place, and reassess for fairness on a regular basis, as the context and inputs change over time. Just as human leaders aren't judged solely on a single good decision, the question of fairness in technology solutions is not a one-time box to check.

Assessing algorithmic fairness

There are many algorithmic fairness toolsets available today. In retail banking, AIB leveraged our in-house fairness toolset to ensure they were ahead of the industry, enabling them to further enhance the integration of algorithmic fairness assessment in the models used to aid their decision making. The fairness toolset helps identify and mitigate bias in algorithms. But what's more, it's designed to facilitate discussions in multi-disciplinary teams and enable relevant action to be taken.

Looking forward: creating more reliable tools for a fairer world

Current machine learning algorithms often focus on correlations (relationships) between variables, rather than causation ("A causes B").

Being able to identify causation would help us make algorithms more interpretable, reliable and fair.

For example, think about a dataset on college admissions with a positive correlation between gender and college admission. Seeing that a higher percentage of males were admitted than females, people might assume a **direct effect**: that gender impacted whether a person was admitted. But what if the majority of women had applied to more competitive departments than the men? In that case, there's a possible **mediator variable**: how competitive the department is in terms of admissions. That could be the cause of the difference in overall admission rates between the genders, rather than gender itself.

“Correlation vs causation” is a long-standing issue that predates sophisticated algorithms or AI systems. But there may be a solution on the horizon for assessing algorithmic fairness, in causal inference.

Accenture Labs is developing a solution for discovering potential mediator variables like the competitiveness of departments in the college admissions example. It uses natural language processing (NLP) to examine existing knowledge bases and uncover previously unknown connections between correlated variables. This can help clarify whether relationships between variables are in fact causal or not.

Enhancement of human skills and expertise

Trustworthy leaders are team players and collaborate well with others. In the context of trustworthy systems, this is about ensuring that humans and technology solutions work together and complement each other.

We worked with leading consumer packaged goods company P&G to apply this approach in formulating better products, faster. Formulation selects, processes and combines a product's ingredients to deliver specific properties, functionality and performance. Products like dishwashing liquid are based on a combination of dozens of ingredients selected from thousands of potential components. The process includes manual steps and can involve physical testing, which can significantly increase the time it takes to get these products to market.

We co-created an AI-based, human + machine toolset that allows P&G formulation developers to amplify their unique talents and knowledge with AI's abilities. It suggests formulations that meet parameters the developer specifies, giving fast, curated inspiration. P&G employees can unleash their creativity in new and unexpected ways, as well as spend more time working on strategic, value-added activities. P&G has a new way to use AI—not just to develop new products but to enable and augment its people.

Human + technology approaches are a powerful driver of value. And when applied thoughtfully, technology can also foster closer collaboration between human colleagues, leading to faster, stronger, and more creative outcomes.

Evaluating uncertainty in technology systems and expanding knowledge

Leaders inspire trust by being well informed. They have relevant experience and understand the larger context around the decisions they're asked to make or influence. As companies work to build and maintain trustworthy technology systems, that means understanding the uncertainty in systems' decisions and outcomes, and designing systems that continue to learn and expand their own bases of knowledge.

Evaluating uncertainty in technology systems

Making good decisions often means questioning things before accepting them as truth. For companies trying to build and maintain trustworthy systems, that means understanding the context around the decisions or outcomes of those systems.

Take algorithms. With every algorithm, there will be some element of uncertainty in its outputs. But companies can show good judgement in their use of these systems by understanding that uncertainty, and taking it into account in how those systems are used. They can even leverage that uncertainty to improve its outputs.

We can use counterfactual explanation systems as an example. By default, all generated counterfactuals will produce the desired output, and for a single point, there are usually many valid counterfactual explanations to "flip the decision." To go back to the example of getting a loan from a bank approved instead of denied, any counterfactual explanation generated would result in the loan being approved. But some of those changes the system proposes might not make sense to humans. That is, they "work" but aren't actually helpful: for example, being told you would have been approved for a loan if the loan amount was negative.

For such a system to be practically useful, it needs to learn how to give more **robust**, **realistic** and **trustworthy** explanations. Working with the Alan Turing Institute, we found a way to improve the outputs of such systems.

By leveraging the predictive uncertainty of the model for which we're generating counterfactual explanations, we can help generate more useful explanations. The model is more confident in an output if the inputs are similar to those it has seen before in the training data. Using the loan scenario from earlier, there would be no examples in the training data of someone applying for a negative amount of borrowed money. Therefore it would be less confident in that output. We use this approach to select the counterfactuals that the model is the most confident in—which are usually the most realistic, useful explanations. This reinforces the "good judgement" aspect of trustworthiness.

Expanding knowledge

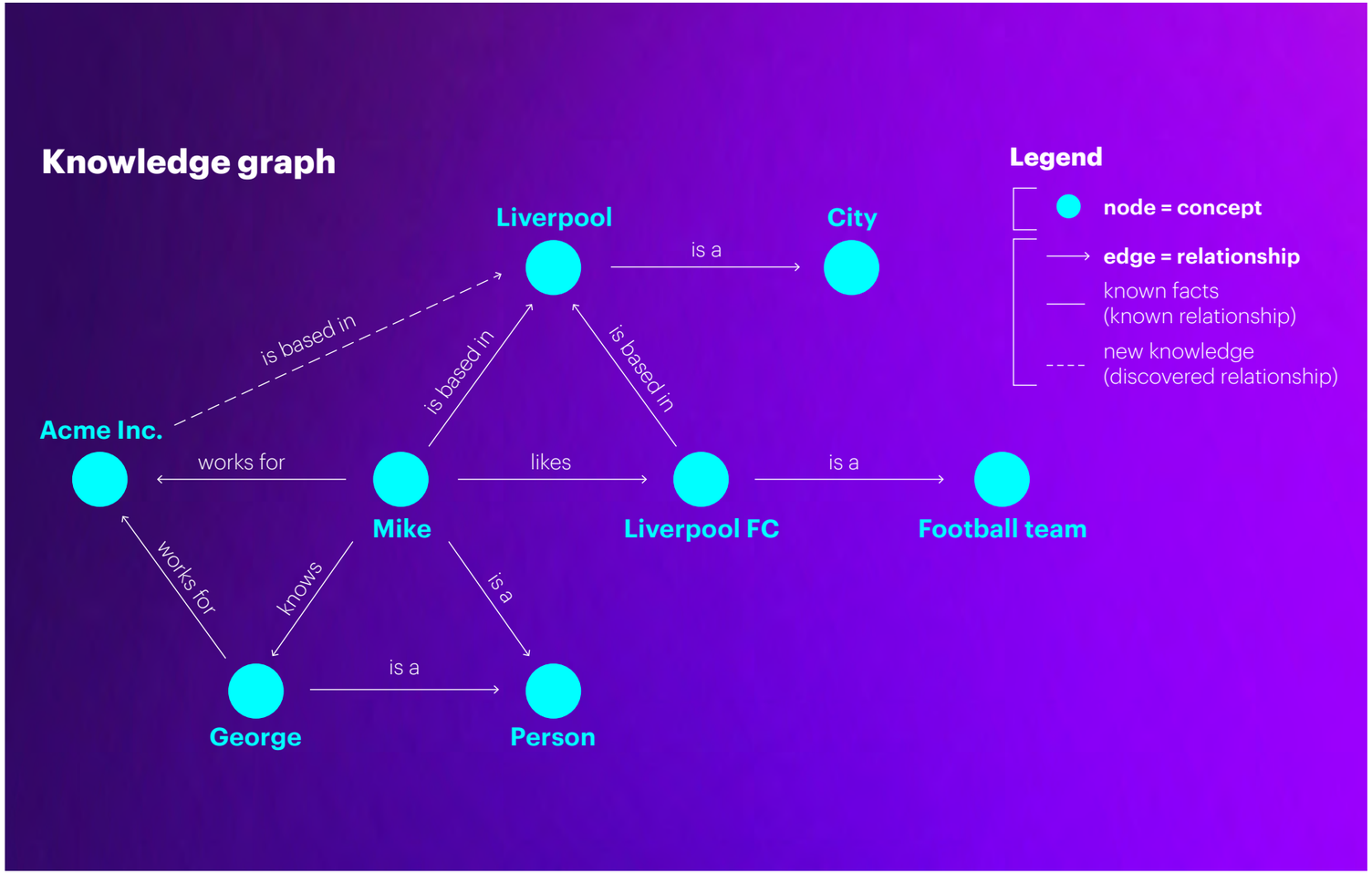
While well-designed technology systems can perform impressively right out of the box, there's always room for improvement.

Just as there's always more for humans to learn, technology systems can be designed to expand both their knowledge base and their own performance.

For example, we've developed a method to apply machine learning to a data representation called knowledge graphs. It discovers new knowledge "hidden" in the data—previously overlooked relationships between pieces of information, or "concepts."

Our solution, Ampligraph, has applications across industries, from targeting new genes in drug discovery to uncovering key relevant new skills for people who are at risk of job disruption. Ampligraph is open source and available at https://docs.ampligraph.org/.

Critically, this method also meets the requirement we discussed earlier around explainability. The system can explain how it discovered these previously-overlooked connections. That's of high importance when it comes to a company deciding to explore a potential connection between a new drug candidate and treatment for a medical condition, or telling an employee to pursue a particular new skill.



03 | Consistency

Robustness, security, privacy and embedding company values

Trustworthy leaders do what they say they are going to do. They consider situations from all angles, so that as little as possible is left up to chance. And they demonstrate their values with their own words and actions. To build trustworthy technology systems, companies must design for the same qualities.

Building robust and secure systems

Trustworthy technology systems must perform as organizations say they will perform, and leave nothing to chance. Amid ongoing security threats, that requires that systems be resilient. They must continue to do what organizations say they will do, even when they're under attack, and they should resist being easily tricked into doing something unexpected or unplanned.

Think about adversarial images. These are pictures that have been deliberately manipulated to trick machine recognition software into misclassifying objects. They can fool machine learning systems even when the object's real identity is obvious to a human. To use a commonly cited example, researchers have manipulated an image of a stop sign so that it isn't recognized by such systems. This was done purely for research purposes, but imagine if someone applied similar techniques to fool the algorithm that a real self-driving car uses to navigate.

Technology systems need to be sufficiently robust to resist this kind of external tampering. Companies can test their AI applications' resilience and robustness to AI-based attacks under a wide range of threat scenarios. A risk framework can help diagnose potential robustness flaws, and an adversarial test suite can probe AI applications for potential flaws and vulnerabilities. By conducting adversarial attacks on a popular object detection model, for example, and then re-training on those attacks, companies can increase the robustness of the model against them.

Resilience at the edge

Smart edge solutions that enable intelligence everywhere will use miniaturized AI-powered models. Unfortunately, by default, those models may not be protected against adversarial AI threats at the same level as their full-scale counterparts. Edge applications are developed for resource-constrained devices with limited power, storage and processing capabilities, which results in compressed ML models (e.g., TinyML). Defense mechanisms developed for trustworthy AI, meanwhile, tend to require a larger and more complex model for protecting them against adversarial AI attacks. Accenture Labs has developed a method of customized pruning of full-scale ML models that not only suits resource-constrained edge devices, but also protects them against security and privacy AI attacks.

Helping ensure data privacy and data security

Human leaders are trusted with important information. Similarly, people must trust technology systems to keep their data private and secure. People increasingly—though often reluctantly—trust their personal data to organizations, even when they don't have direct relationships with the companies in question.

Meeting expectations around privacy and security means that trustworthy systems only use data collected from end-users in transparent, agreed-upon ways. In combination with this type of transparency, strong governance and regular policy reviews, though, organizations also have various technology-driven options for keeping data private and secure. There are multiple techniques that can be used or combined to safeguard information. Some organizations use differential privacy, where data about patterns within a dataset is publicly available, but data on individuals is withheld. Another option is using synthetically generated data, rather than real data, to train machine-learning models.

Synthetic datasets mimic the statistical properties of the original data without containing the original data. In an ideal scenario, the resulting dataset is close enough to be useful for analysis, but different enough to protect the privacy of the individuals involved. But finding that balance takes effort. Tools like the Accenture Automated Privacy Assessment Tool can help evaluate synthetic data with respect to privacy, utility and similarity. For example, the tool evaluates the privacy level of synthetic datasets by quantifying re-identification risk—that is, whether it's possible to identify the individuals whose data the set was originally based on—while ensuring that the dataset is still valuable for larger analysis.

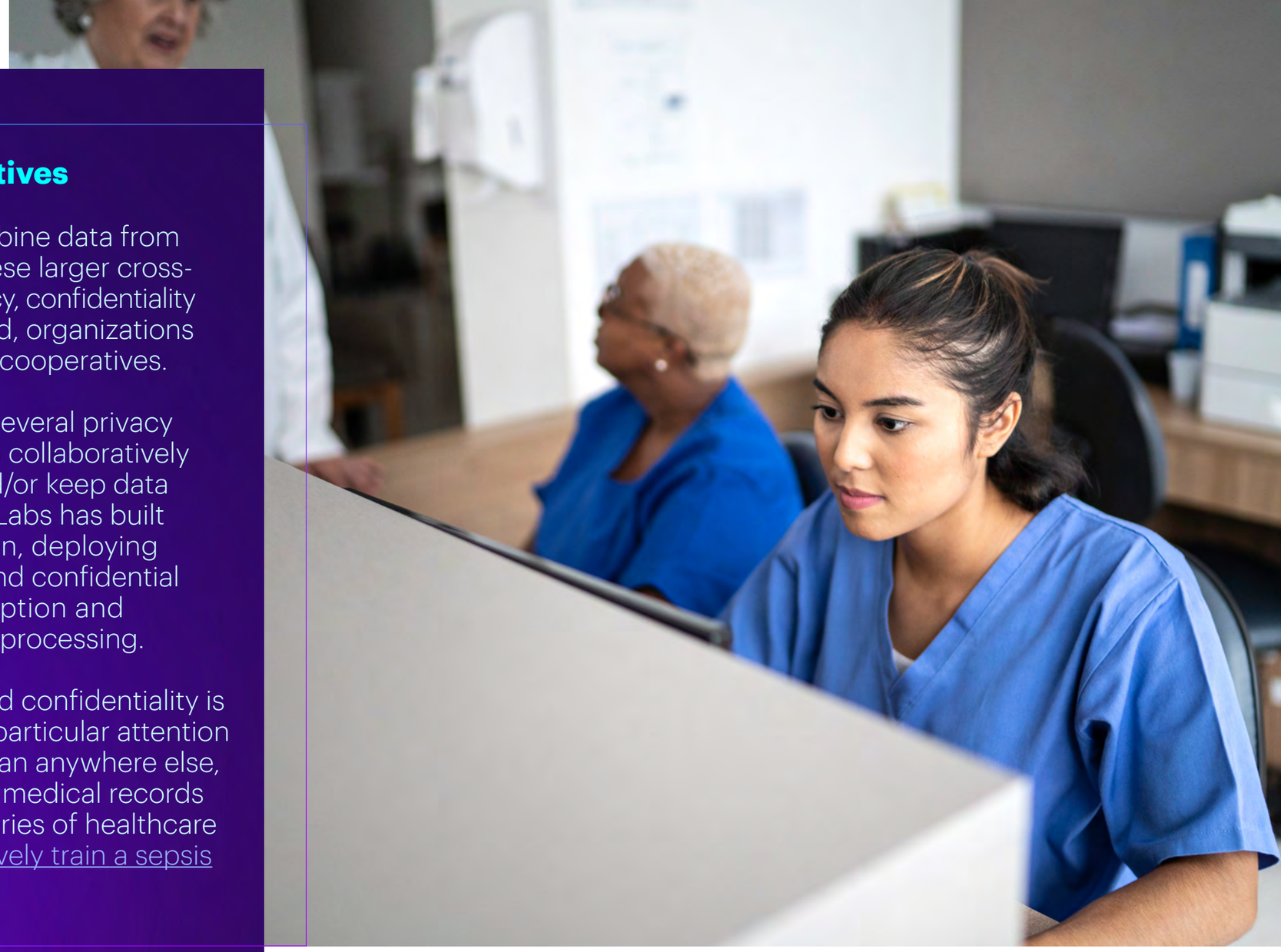
This type of data synthesis approach is being applied to electronic health records; the goal is to allow for the sharing of realistic synthetic medical data in a way that helps enable value and insights from the original data to be extracted while **preserving patients' privacy**.

Advancing privacy with data cooperatives

Organizations are increasingly looking to combine data from disparate sources to drive value. To enable these larger cross-group collaborations while preserving the privacy, confidentiality and ownership of the information being shared, organizations are turning to the emerging approach of data cooperatives.

These cooperatives are based on combining several privacy preserving techniques (PPTs); participants can collaboratively work on data without central aggregation, and/or keep data encrypted while being processed. Accenture Labs has built a privacy-preserving data cooperative solution, deploying several PPTs that include federated learning and confidential computing, multi-party homomorphic encryption and computation and privacy preserving data pre-processing.

Data collaboration with this level of privacy and confidentiality is valuable in many industries, but has received particular attention in healthcare. In this industry perhaps more than anywhere else, maintaining data privacy is essential: patients' medical records must be protected. Our solution supports a series of healthcare related scenarios, and was used to collaboratively train a sepsis detection model across hospitals.

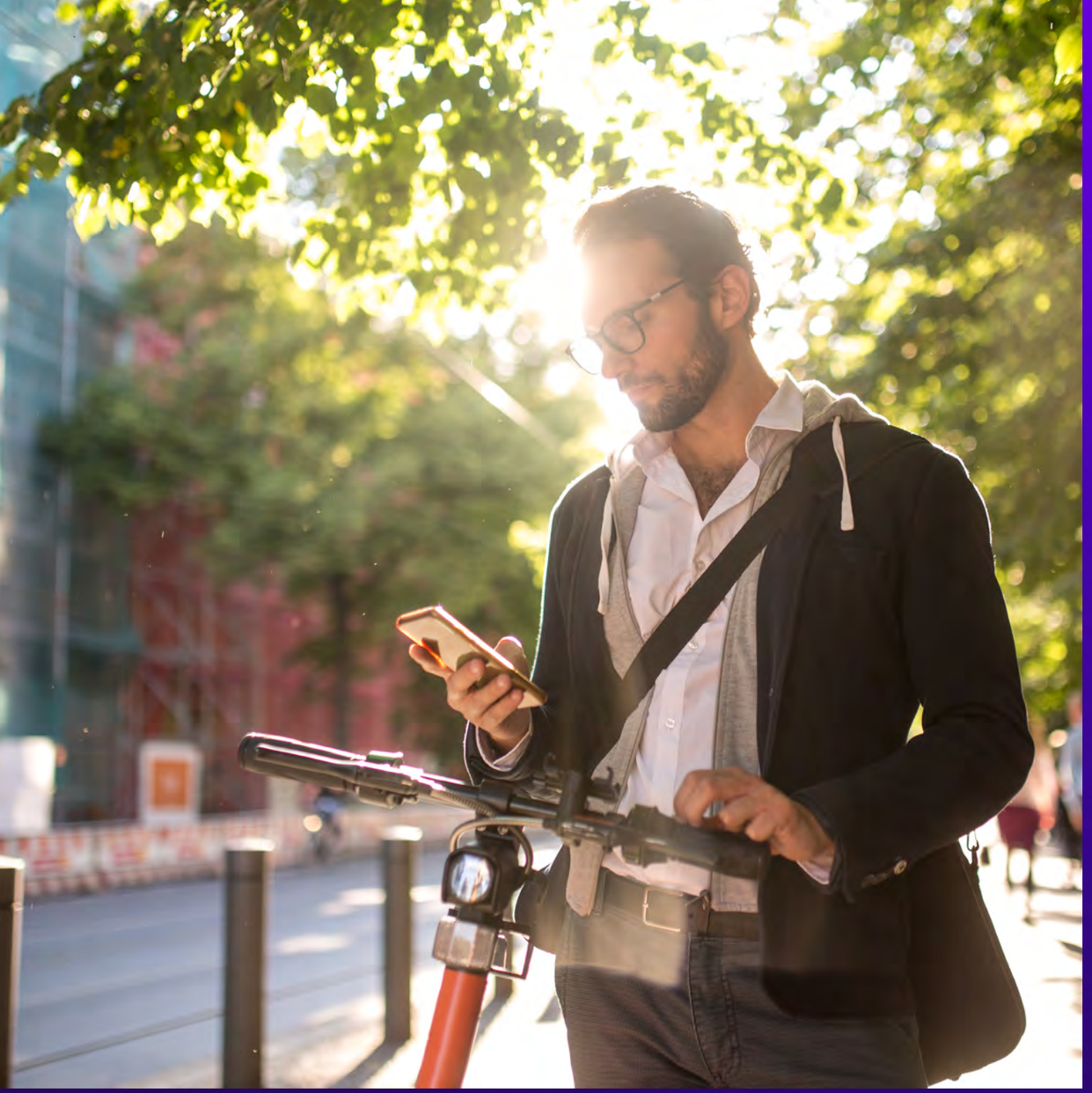


Embedding company values

Trustworthy leaders live their values. Companies can similarly use trustworthy technology systems to reflect and reinforce their own values at scale.

For example, organizations can ensure that models support their diversity goals by making the relevant decisions in the algorithm design. It's a powerful way to automate company values, by embedding principles into machine code.

Of course, accountability is key. Leaders in trustworthy systems set quantitative goals that reflect their values. For example, many companies say they are customer-first. But how many commit to, and are accountable for, passing on a defined percentage of savings from automation back to the customer? Accenture has worked with Veritas, an industry consortium established by the Monetary Authority of Singapore to define a framework that helps enable organizations to translate their values into quantitative commitments with the measurement and accountability of these commitments incorporated into their processes.

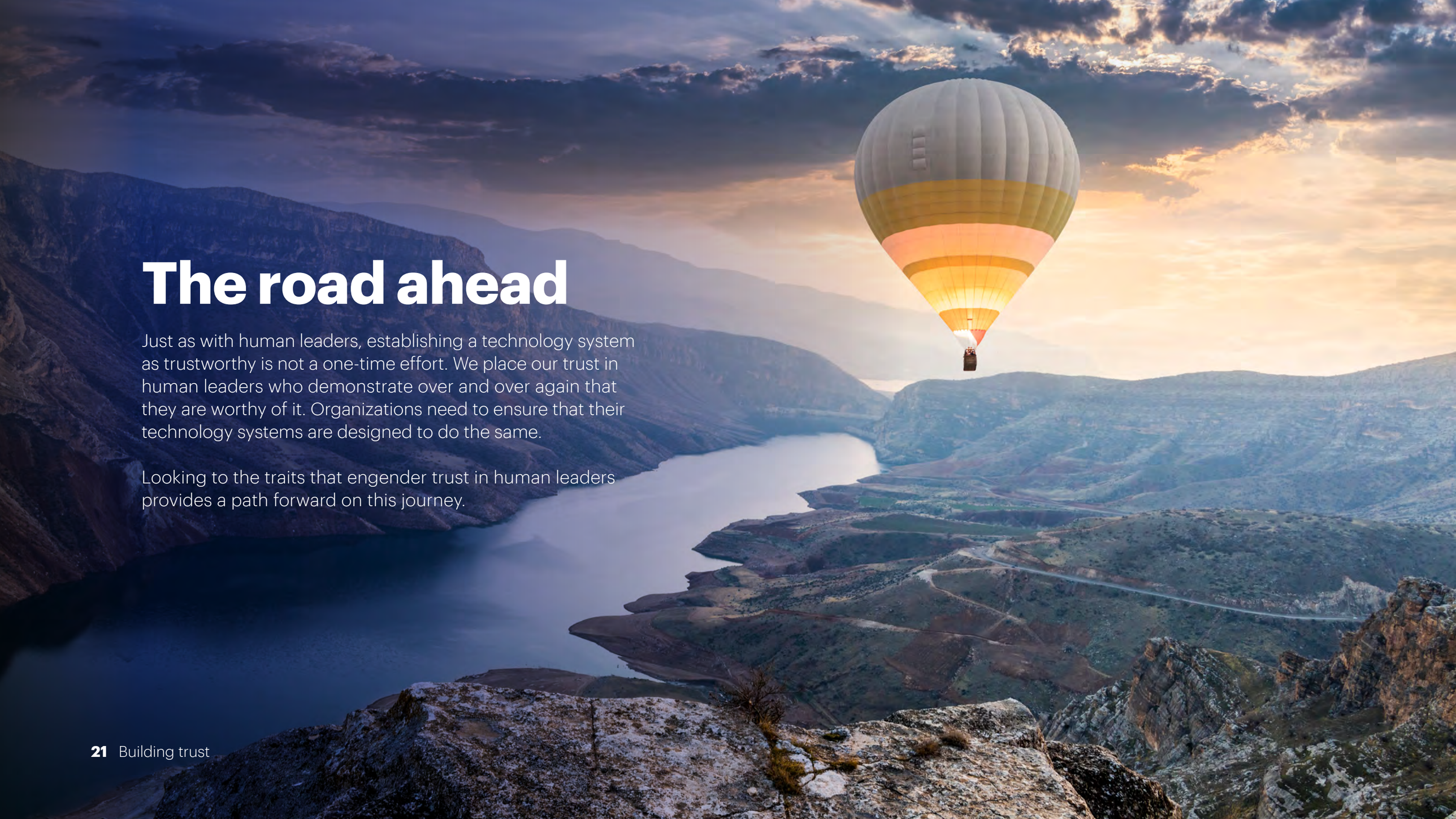


Building more environmentally sustainable technology

Leaders are looking to enable inclusive growth and sustainable development, looking for ways to reduce their environmental impact—including that of their technology systems.

For example, training highly complex AI models often requires a staggering level of energy consumption. Accenture Labs has developed a technical toolset that highlights the implications of machine-learning design, development, and testing choices on energy efficiency and sustainability. It helps organizations develop energy-efficient machine learning and find the right balance between building a reliable model and mitigating the environmental impact.

Emerging hardware solutions will also play a role in enabling heterogenous compute, where organizations can apply different types of hardware to different tasks and optimize for efficiency and power savings. For example, traditional computing architectures need a lot of power to perform machine learning tasks. Alternative architecture approaches like neuromorphic computing can provide low-power intelligence at the edge. As methods like these reach maturity, they'll provide another path toward energy-efficient systems.



The road ahead

Just as with human leaders, establishing a technology system as trustworthy is not a one-time effort. We place our trust in human leaders who demonstrate over and over again that they are worthy of it. Organizations need to ensure that their technology systems are designed to do the same.

Looking to the traits that engender trust in human leaders provides a path forward on this journey.

Trustworthy human leaders maintain positive relationships. For organizations looking to establish trustworthy technology systems, that means providing clear explanations about how decisions are made; assessing and addressing the fairness of the technology solutions they use; and finding the right balance between humans and technology solutions.

To establish and maintain the good judgement and expertise demonstrated by human leaders in technology systems, organizations must understand and take into account the uncertainty around the decisions and outcomes of those systems. They must also design systems to continue to learn and expand their knowledge bases.

Finally, organizations should strive to emulate the consistency of trusted human leaders with their technology systems. They should ensure that their systems perform as intended even when under attack—that they show resilience and robustness, like a trusted human leader in a time of stress. They must keep their commitments to safeguard information, maintaining data privacy and data security; and ensure that the core values of their organization are reflected in their use of technology systems, with quantitative commitments that go beyond the minimum to exceed expectations, in measurable, accountable ways.

Many of these traits of trustworthy leaders may sound straightforward, but acquiring and maintaining them isn't easy for humans or technology systems. It requires sustained effort. Building systems that incorporate fairness, transparency, robustness and explainability is just the first step. Organizations also need the right governance, security and culture to bring trustworthy systems to life. **As technology continues to advance, new challenges and potential pitfalls are emerging all the time. Standing still isn't an option. But for those who act, there's a big opportunity.**

Companies that understand and commit to building trustworthy systems stand to increase market share while reducing churn. They can automate more of their processes and boost efficiency, without compromising employee or customer experiences. They can form new partnerships to extend their reach. They can maintain compliance with evolving regulations. The same characteristics that inspire trust in leaders can help organizations build trust in their systems. **Those organizations will become trusted too—and when they innovate, people will follow**.

References

1 "The 3 Elements of Trust." Jack Zenger, Joseph Folkman. Harvard Business Review. https://hbr.org/2019/02/the-3-elements-of-trust

2 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, NeurIPS, 2018. R. Mc Grath, L. Costabello, C. Le Van, P. Sweeney, F. Kamiab, Z. Shen, F. Lecue Interpretable Credit Application Predictions With Counterfactual Explanations.

3 Data Cooperative is used to collaboratively train a Sepsis detection model.

Contacts

Medb Corcoran
Managing Director, Accenture Labs
Global Responsible AI Lead, Accenture Tech Innovation
medb.corcoran@accenture.com

Ray Eitel-Porter
Managing Director, Applied Intelligence
Global Lead, Responsible AI
ray.eitel-porter@accenture.com

Contributors

Luca Costabello, Louis DiValentin, Giuseppe Giordano, Amin Hassanzadeh, Jer Hayes, Laetitia Kameni, Rory M. Mc Grath, Mohamad Nasr-Azadani, Bogdan E. Sacaleanu, Shubhashis Sengupta, Steven Tiell

About Accenture

Accenture is a global professional services company with leading capabilities in digital, cloud and security. Combining unmatched experience and specialized skills across more than 40 industries, we offer Strategy and Consulting, Technology and Operations services and Accenture Song—all powered by the world's largest network of Advanced Technology and Intelligent Operations centers. Our 710,000 people deliver on the promise of technology and human ingenuity every day, serving clients in more than 120 countries. We embrace the power of change to create value and shared success for our clients, people, shareholders, partners and communities. Visit us at www.accenture.com.