Customers who trust products and services engage more, which improves their experience and satisfaction. AI tools now shape many customer decisions, so their trustworthiness is critically important. Jennifer Shkabatur and Alex Mintz explore how commonly accepted measures of trustworthiness in AI can be practically tested and ranked.

Developing a Trustworthy AI Rating System and Its Impact on Customer Engagement

Jennifer Shkabatur
Reichman University

Alex Mintz
Reichman University

Every day artificial intelligence (AI) products shape a broad array of customer decisions about finance, e-commerce, health care, leisure, professional recommendations, and more. Indeed, many businesses have drastically changed their modes of operation, deploying AI so they can better understand their customer's preferences, shape their decisions and behavior, and strengthen their engagement.[1]

Since customers' trust in products drives their engagement, the question of whether AI products can be trusted is critically important.

Since customers' trust in products drives their engagement, the question of whether AI products can

be trusted is critically important. We have devised a method by which commonly accepted measures of trustworthy AI can be practically tested, and a ranking scale for trustworthiness in various sectors. This method can be applied by the business community, investors, regulators, ranking agencies, and the customers themselves.

Trustworthiness as a driver of customer engagement

A longstanding and primary task of marketers is to persuade consumers who are on the fence about a product or service to go ahead and buy.[2] In order to achieve this goal, they must build trust — confidence in the product's reliability, robustness, security, integrity, and ability to meet their needs.[3]

Trust is therefore generally recognized as one of the primary factors that strengthen customer engagement, sales, and satisfaction.[4]

Trust contributes to a customer's willingness to purchase the product again, builds a deeper and more personal connection with the product, and thus increases sales[5] and, ultimately, the probability of future repurchases.[6] Conversely, lack of trust in a product reduces engagement in both individual and business customers.[7]

To increase customer engagement in all interactions, businesses use a range of AI products, including recommendation systems, conversational agents, sentiment analyzers, and natural language processing algorithms. While customers draw huge benefits from these AI products, they may also find assessing their trustworthiness to be challenging for several reasons.

1. The opacity of AI operations tends to prevent public scrutiny, particularly since AI systems are difficult to explain and understand.[8] Customers may respond to this uncertainty with bias and discrimination that

Table 1. Assessment Questions for Trustworthy AI

1. Human agency and oversight	• *Self-assessment.* Does the AI product allow the user to reasonably assess or challenge the product? • *Intervention.* Does the AI product enable human intervention during its decision cycles? • *Discretion.* Is it possible to integrate human discretion into the operation of the AI product? • *Override.* Is it possible for a human to override a decision made by the product?
2. Technical robustness and safety	• *Effectiveness.* Does the AI product achieve the outcomes that it promises to achieve? • *Accuracy.* Does the AI product make correct predictions, recommendations, or decisions? • *Safety.* Is the AI product safe to use (in its data security protocols, and software and hardware safety)? • *Reproducibility.* Does the AI product exhibit the same behavior when applied under similar conditions?
3. Privacy, data governance, and legal compliance	• *Data sources.* What are the data sources used by the AI product? How and when was this data collected? Was consent acquired? How is this data used by the AI product? How is the integrity of the data ensured? • *Data access and management.* Who has access to the data utilized by the AI product, and under what conditions? Who is eligible to change or manage the data? • *Legal compliance.* Does the AI product comply with the legal requirements of the country in which it is deployed (e.g., privacy, health and safety, etc.)?
4. Transparency and explicability	• *Traceability.* Are the data and processes that yield the AI product's decision documented? Is it possible to trace back and link the product's inputs and outputs? • *Explicability.* Is it possible to explain in plain language how the AI product works? • *Communication.* Is any information provided regarding the limitations of the AI product, its level of accuracy, or other issues related to its operation?
5. Diversity, non-discrimination, fairness	• *Biases.* Can the data used by the AI product be affected by inadvertent bias, discriminatory patterns against certain groups, or incompleteness? Can the outcomes of the AI product lead to discrimination against certain groups? • *Accessibility.* Is the AI product user-centric and designed in a way that allows anyone to use it, regardless of age, gender, abilities, or characteristics?
6. Societal and environmental well-being	• *Sustainability.* What are the environmental impacts of the AI product (including its development, deployment, and usage)? What are the resource usage and energy consumption during the product's training and operation? • *Social impacts.* Can the AI product adversely affect users' mental or social well-being?
7. Accountability	• *Auditability.* How practical is it to audit the AI product? Can it be independently audited? • *Minimizing and reporting negative impacts.* How can the AI product's actions or decisions be reported? • *Redress.* If the AI product causes adverse effects, what redress venues are available?

Figure 1. Assessing Three AI Products

We analyzed and rated for trustworthiness the following products:[15]

1) **Fitness coaching app** that offers personalized coaching for improving fitness and mindfulness, as well as developing healthy lifestyle habits. Users indicate the goals they want to achieve and answer a survey about their habits. With this information, the app generates a personalized training plan for each user.
2) **AI-backed search engine** that analyzes various types of financial documents, relies on natural language processing for improved results, and transforms unstructured data into structured. It then allows users to search for specific terms within the documents based on geography and time. It includes features such as 'synonyms search,' and sentiment analysis, and prioritizes results according to importance and relevance.
3) **Conversational AI-based virtual assistant** that specializes in banking. It helps users analyze their financial activities, generating information and recommendations based on various factors and prior interactions.

companies struggle to reverse or even uncover.

2. Low public awareness and AI literacy, which may lead the public to neglect and disregard the risks associated with AI.[9] A recent survey showed that only 62 percent of respondents have any knowledge of AI and the majority reported that they had low understanding of AI. When presented with a range of common AI applications, respondents were not aware that the described technology used AI.[10]
3. Rapid scaled deployment of AI that hinders companies in properly scrutinizing AI products before investing in and deploying them.[11]
4. Lagging regulation of AI systems such that governments cannot provide timely and adequate legal recourse when AI problems arise, complicated by differences in the rules of separate countries.[12]

A vast range of stakeholders, including private companies, academics, government agencies, intergovernmental organizations, and professional associations, have formulated principles for ethical, trustworthy, and human-centered AI systems that customers could trust. By examining these principles, we traced significant commonalities and emerging norms.

In 2019, the European Commission (EC) proposed a seven-principle framework for 'trustworthy AI' systems that has been broadly accepted:[13]

1. **Human agency and oversight:** AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights.
2. **Technical robustness and safety:** AI systems need to be resilient and secure, offering their users accurate, reliable, and effective services.
3. **Privacy and data governance:** Full respect for privacy and data protection should be ensured.
4. **Transparency and explicability:** AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.
5. **Diversity, non-discrimination, and fairness:** Unfair bias and outputs that discriminate against specific groups must be avoided.
6. **Societal and environmental well-being:** AI systems should benefit all human beings, including future generations. They must be environmentally friendly and sustainable.
7. **Accountability**: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.

Many subsequent AI ethics frameworks focus and draw upon these principles.[14] Still, while we have achieved notable unity as to the characteristics of trustworthy AI products, few governments or organizations use those characteristics to actually assess AI products. We have therefore developed a procedure for testing AI products against the EC's parameters.

Assessing the trustworthiness of AI products

We begin by translating these parameters into questions which we then use to assess three real AI products in common areas of customer engagement: finance, health and fitness, and natural language processing. The resulting ratings offer a standardized assessment of trustworthiness across AI products which can be used by consumers, regulators, businesses, and investors and which therefore can be a significant driver of customer engagement.

We measure the trustworthiness of AI products based on the questions in Table 1. These questions reflect the EC's definitions and also closely follow the wording used in other studies of trustworthy or ethical AI.

This straightforward scale offers businesses and consumers an easy, standardized, and convenient way to assess the trustworthiness of an AI product.

Table 2. Trustworthiness Assessment and Rating

	Personalized coaching app	AI-backed search engine	Conversational AI-based virtual assistant
Human agency and oversight	The coaching plan's objectives are determined by the user. Changing one's preferences results in changed outputs.	The outputs are modified interactively in response to user changes. But the user cannot challenge the system or establish discretion as part of its operation.	The chatbot is interactive. If the bot cannot understand the user's inputs, it suggests chatting with a live agent.
Technical robustness and safety	The product reflects the user's preferences, yet it does not claim or prove that its recommendations are optimal. It exhibits similar behavior when repeated under similar conditions.	The system displays the data that it finds, so accuracy is easy to establish. However, there may be missing results, and no convenient way to account for these.	The system does not undertake specific actions but rather provides recommendations. There is no information on whether these are optimal.
Privacy, data governance, and legal compliance	The privacy policy is very clear, complying with the laws of California. It states what, how, and when data is collected; requires user's consent for use of the data; and explains under what conditions data can be shared with third parties.	The system's big data is not related to its users. It accesses extremely limited user information (e.g., email and name), and is highly secured.	The privacy policy elaborates on user data usage. However, while user data is used to improve the service, it is not explained how data integrity is maintained.
Transparency and explicability	The system's outputs are explained, but does not explain why these outputs were preferred over others. There is no information regarding limitations, and no way to trace back to its operations mode.	The system highlights the keywords and search terms that it found, yet there is no explanation of why and how certain search terms are prioritized. The sentiment analysis is also not explained.	The bot responds as part of a conversation, but there is no explanation of how financial recommendations are made.
Diversity, non-discrimination, fairness	The outputs are based on user-specified goals and body characteristics. However, it is not clear which biases may emerge against certain user groups.	Some AI features do not support non-English content. The system focuses on financial documents and there seems to be no discrimination against certain user groups, yet there is no clear way to test this.	No sufficient information for a response. Some user groups may be discriminated against by the virtual assistant. There is no way to compare financial recommendations provided to different users.
Societal and environmental well-being	No information regarding environmental impacts. Since it is a fitness app, it is expected to positively affect users' well-being.	The system uses Amazon Data Centers. Its general societal impact seems to be neutral.	No information, but the system does not appear to raise considerable concerns.
Accountability	Users can change the suggested workout plan. They can also contact the company's support desk. It is unclear whether and how redress for adverse outcomes could be provided.	There is no practical way to audit the system, but users can contact the company's support desk. It is unclear whether and how redress for adverse outcomes could be provided.	There is no practical way to audit, but users can contact the company's support desk. It is unclear whether and how redress for adverse outcomes could be provided.

We used these questions to assess the trustworthiness of real AI products and developed a trustworthy AI scale ranging from green (full compliance with the principles), to yellow (partial compliance), to red (lack of compliance and/or impossible to check). This straightforward scale offers businesses and consumers an easy, standardized, and convenient way to assess the trustworthiness of an AI product and, as a result, its impact on customer engagement.

We pilot tested the proposed method on three real AI products, which were randomly selected from a database of 1,700 that we assembled from key sectors that use AI. Dr. Shkabatur and an independent, highly skilled computer scientist conducted the test by examining the products' websites and running demos. Using the questions in Table 1, we rated the products as green, yellow, or red.

Figure 1 provides general information about the AI systems. Their trustworthiness ratings appear in Table 2.

This pilot examination of our rating method gives a sense of whether the trustworthy AI parameters could be applied to real AI products. It has led us to the following reflections:

1) Much debate has been dedicated to privacy and data governance concerns about AI products. Our analysis suggests that the designers of these products have a considerable awareness of privacy requirements and expectations, and that they could feasibly test for and achieve compliance with privacy protection guidelines. Assessing technical robustness also does not appear, by this initial analysis, to present particular difficulties for designers of AI products that target general customer markets.
2) It is possible to examine whether the algorithms in AI products are discriminatory and produce biased results.
3) We found it tougher to assess trustworthiness principles that require peering into the AI products' black box. Transparency and explicability are more difficult for products that operate on more complex datasets. Accountability presents a challenge, because AI products typically do not provide information on whether they could be externally audited, or how negative impacts could be redressed. This generates a particular challenge for customer engagement since AI systems do not typically provide sufficient information on whether customers can trust them to achieve fair and transparent outcomes.
4) Assessing the societal and environmental well-being of these AI products is not easy, raising hurdles for ethical customer engagement. Generally, neither consumers nor businesses know or have the tools to assess how the AI system affects users. Measuring those affects would again require opening the black box.

A final word

Given the widespread use of AI products for customer engagement and the growing reliance on them among businesses and customers, the trustworthiness of these products must be thoroughly scrutinized by investors, businesses, regulators, and the users themselves. Indeed, the question of whether AI products can be trusted is critical and has received considerable attention.[16]

Our procedure for rating the trustworthiness of AI products is rooted in a practical application of commonly accepted EC parameters.

Our procedure for rating the trustworthiness of AI products is rooted in a practical application of commonly accepted EC parameters. It measures the trustworthiness of AI products on a straightforward scale of green-yellow-red, which could be readily understood by customers. This scale could work in a way similar to the hygiene ratings that restaurants display in their front windows. An impartial authority such as a ranking agency, academic body, or the AI company itself could administer the ranking process, provided it offered a full explanation on each parameter.

While an external examination can effectively assess certain common parameters of trustworthy AI, others require opening the black box, an endeavor that will require further consideration. We suggest that only products graded green across all parameters should be considered fully trustworthy, while those with some yellow grades should be viewed as partially trustworthy.

Acknowledgement

The authors thank Eran Hadas and Shira Akiva for their research assistance. ■

Author Bios

Jennifer Shkabatur is an Assistant Professor at the Lauder School of Government, Diplomacy & Strategy and Senior Researcher at the Computerized Decision-Making Lab at Reichman University in Israel. Her research focuses on the ethics of artificial intelligence, innovation in the digital economy, and data and information policies. She also consults on these issues with the World Bank, UNDP, ILO, and more. She earned doctorate and master's degrees from Harvard Law School. jshkabatur@runi.ac.il

Alex Mintz, is Director of the Computerized Decision Making Lab at Reichman University. He was awarded the Lifetime Achievement Award from the Israeli Political Science Association in 2019 and has received numerous other accolades. Mintz was editor in chief of the journal *Political Psychology* and on numerous editorial boards. Among his many books are: *The Polythink Syndrome: U.S. Foreign Policy Decisions on 9/11, Afghanistan, Iraq, Syria, Iran,* and *ISIS,* and *Understanding Foreign Policy Decision Making.*

Endnotes

1. See:
   Kumar, V., Rajan, B., Venkatesan, R., & Lecinski, J. (2019). "Understanding the role of artificial intelligence in personalized engagement marketing," *California Management Review*, 61(4):135–155.
   Davenport T., Guha A., Grewal D., Bressgott T. (2020). "How artificial intelligence will change the future of marketing," Journal of the Academy of Marketing Science, 48 (1):24-42.
   Bag, S., Srivastava, G., Al Bashir, M. M., Kumari, S., Giannakis, M., & Chowdhury, A. H. (2021). "Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion." *Benchmarking*.
2. See:
   Morgan, R.M., and Hunt, S.D. (1994). "The commitment-trust theory of relationship marketing," *Journal of Marketing* 58:20–38.
   Sichtmann, C. (2007). "An analysis of antecedents and consequences of trust in a corporate brand," *European Journal of Marketing* 41(9/10):999–1015.
3. Delgado-Ballester, E., Munuera-Aleman J.L., Yague-Guillen, M.J. (2003). "Development and validation of a trust scale," *Int'l J. Market Research* 45(1):35-58.
4. See:
   Choung, H., David, P., & Ross, A. (2022). "Trust in AI and its role in the acceptance of AI technologies." *Int'l Jour. Human–Computer Interaction*, 1–13.
   Ferrario, A., Loi, M., Vigano, E. (2019). "AI we trust incrementally: A multi-layer model of trust to analyze human-artificial intelligence interactions," *Philosophy & Technology*, 1-17.
   Wang, X., Tajvidi, M., Lin, X., Hajli, N. (2019). "Towards an ethical and trustworthy social commerce community for brand value co-creation: A trust-commitment perspective." *J. Bus. Ethics*, 1-16.
   Shin, D. (2020). "User Perceptions of Algorithmic Decisions in the Personalized AI System: Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability," *Journal of Broadcasting & Electronic Media*, 64(4):541-565.
5. See:
   Bowden, J. L-H. (2009). "The Process of Customer Engagement: A Conceptual Framework," Journal of Marketing Theory and Practice, 17:1, 63-74.
   Singh, J.J., Iglesias, O. & Batista-Foguet, J.M. (2012). "Does Having an Ethical Brand Matter? The Influence of Consumer Perceived Ethicality on Trust, Affect and Loyalty," *J. Bus. Ethics*. 111:541–549.
6. Madhani, P. M. (2020), "Ethics in Sales and Marketing: Key Advantages", *Marketing Mastermind*, 17(5):53-58.
7. See:
   Kingshott, R.P., Sharma, P., Chung, H.F. (2018). "The impact of relational versus technological resources on e-loyalty," *Industrial Marketing Management*, 72:48-58.
   Grewal, D., Guha, A., Satornino, C.B. and Schweiger, E.B. (2021), "Artificial intelligence: the light and the darkness", *J. Bus. Research*, 136:229-236.
8. See:
   Rudin, C. (2019), "Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead," *Nat. Mach. Intell.* 1:206–215.
   Rai, A. (2020), "Explainable AI: from black box to glass box," *J. Acad. Mark. Sci*. 48:137–141.
   Guha, A., Grewal, D., et al. (2021). "How artificial intelligence will affect the future of retailing." *Journal of Retailing*, 97(1):28-41
9. Curtis, C., Gillespie, N. & Lockey, S. (2022). AI-deploying organizations are key to addressing 'perfect storm' of AI risks. *AI Ethics*.
10. Gillespie, N., Lockey, S., Curtis, C. (2021). Trust in Artificial Intelligence: a five-country study. *The University of Queensland and KPMG Australia*.
11. Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., Shoham, Y., Clark, J., Perrault, R. (2021). "*The AI Index 2021 Annual Report*".
12. See:
    Mittelstadt, B. (2019), "Principles alone cannot guarantee ethical AI," *Nat. Mach. Intell.* 1:501–507.
    Taeihagh, A. (2021). "Governance of artificial intelligence." *Policy and Society*, 40(2):137-157.
    Djeffal, C., Siewert M. B., Wurster S. (2022). "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies," *Journal of European Public Policy*.
13. These principles rely on four core ethical principles that lie in the basis of the EU Charter: respect for human autonomy; prevention of harm; fairness; and explicability.
    European Commission (2019). "Ethics guidelines for trustworthy AI." *EC HLEG* https://digital-strategy-ec-europa-eu.ezprimo1.idc.ac.il/en/library/ethics-guidelines-trustworthy-ai.
14. The OECD developed a typology, which includes roughly similar principles ("Inclusive growth, sustainable development and well-being;" "Human-centred values and fairness;" "Transparency and explainability;" "Robustness, security and safety;" and "Accountability.").
    OECD/LEGAL/0449. (2019). "Recommendation of the Council on Artificial Intelligence." OECD.
    The IEEE's Ethically Aligned Design for Autonomous and Intelligent Systems resulted in comparable principles.
    IEEE (2019) "Ethically Aligned Design – First Edition", *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*.
    See also: Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches*., which compares 36 distinct initiatives to define "ethical" and "rights-based" AI and identifying largely similar themes.
15. The products were tested in July 2022.
16. See:
    Ameen, N., Tarhini, A., Reppel, A., & Anand, A. (2021). "Customer experiences in the age of artificial intelligence." *Comp. Human Behavior*, 114:106548.
    Dwivedi, Y.K, Hughes L., et al. (2019). "Artificial Intelligence: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy." *Int'l Jour. Info. Management*.