Solving the Trust Problem: Steps to Strengthen Confidence in AI-Driven Business Applications

A Research Paper submitted to the Department of Engineering and Society.

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Irvine Madenga

Spring, 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Bryn E. Seabrook, Department of Engineering and Society

"Solving the Trust Problem: Steps to Strengthen Confidence in AI-Driven Business Applications”

Introduction:

On January 27th, 2025, a major panic spread across the tech industry and stock market. News outlets and social media all over the web were astonished as AI leader Nvidia’s market cap dropped by $593 billion in a single day, the highest one-day loss for any company in the history of Wall Street (Carew et al., 2025). Even tech giants such as Microsoft and Google’s parent company, Alphabet, had their shares fall by 2.1% and 4.2%, respectively. Now, what caused all this mayhem, you may ask? On January 20th, 2025, a Chinese AI company called Deepseek released its newest model, Deepseek R1. The Deepseek chatbot is very similar to OpenAI’s Chat GPT and shot up to the top of app stores, quickly overtaking Chat GPT. Several elements of the development of this model greatly differentiated it from its competitors and caused severe concern among tech giants in the United States. Firstly, its performance is on par with OpenAI’s o1 model that was released at the end of 2024, however, it was produced at a fraction of the price. Reports claim that only $6 million was spent by Deep Seek in the development of the model, in contrast to the over $100 million spent by Open AI to produce their models, and despite restrictions that the American government put on China from receiving powerful electric chips. Additionally, the platform is completely free to use with no limits (unlike Chat GPT, which has limits to free usage and has a paywall behind stronger models) and is also open source, meaning that the entire codebase is available for anyone to see with no barrier. (*What Is DeepSeek - and Why Is Everyone Talking about It?*, 2025) This led many to question the future of artificial intelligence in society and how much it will change based on several unpredictable factors. The effect that the emergence of Deep Seek had on the stock market and tech industry shows the prevalence and volatility of artificial intelligence in the current world. The magnitude of the effect that AI development has had on businesses and economies globally is already immense, but let us also consider its consequences more conceptually, with more individual interactions and use.

Companies in all sectors, whether in healthcare, finance, manufacturing, or retail, have been leveraging AI to improve business processes. From market analysis to drug development, the wide range of use cases that AI possesses has caused it to permeate every part of society. With the aforementioned pervasiveness of AI in current society comes several concerns when it comes to how secure one feels about its use in things that affect their daily lives and benefit from their information. Additionally, the speed at which AI has been developing in recent years provokes uncertainty and doubt about the development of these systems. With little policy and legislation to regulate artificial intelligence at the current moment, a lot of the power of these systems lies with the AI companies producing these technologies. Artificial intelligence is also unique in that much of its functionality is automated, and machine learning algorithms are capable of training themselves and improving their functionality with data that is fed to them. With this in mind, consideration should be given to how much agency these autonomous systems have, as well as how much one can be assured that those working on these technologies behind the scenes are taking measures to consider this. How can corporate entities improve trust in AI-driven business/systems?

Methods:

In the context of improving trust in AI-driven business applications, understanding the interplay between human and machine agency within human-machine networks (HMNs) is crucial. As machines increasingly take on active roles in these networks, they influence human actors and reshape trust dynamics. Engen, Pickering, and Walland (2016) highlight that the active participation of machines not only facilitates interaction but also necessitates a nuanced understanding of trust within socio-technical systems. Bankins (2021) further underscores the importance of ethical considerations in AI deployment, particularly in human resource management, advocating for frameworks that balance human and machine agency to foster trust. Trust in AI technologies is shaped not just by their capabilities but also by users' perceptions of reliability and agency, as noted by Choung et al. (2023). As machines gain autonomy, understanding their role becomes vital for enhancing collaboration and trust. To maximize the benefits of AI while mitigating potential harms, ongoing research should explore the relationship between machine agency and trust, integrating ethical guidelines into the design of AI systems (Li et al., 2023; Mylrea & Robinson, 2023; Oyeniran et al., 2022). Addressing these factors will be essential as organizations seek to implement AI technologies that are both effective and trustworthy, ultimately fostering a more positive relationship between humans and intelligent systems. The breadth of sources-including academic journals, news articles, and books-reflects the interdisciplinary nature of research on trust in AI. This diversity ensures that the discussion is informed by both theoretical frameworks and practical perspectives, capturing the evolving challenges and opportunities in building trustworthy AI systems across different domains.

Background:

To understand the effects of AI and issues regarding trust and agency, one must understand what AI is and how it’s used. Artificial intelligence, as described by Stanford Professor John McCarthy in 1955, is “the science and engineering of making intelligent machines”. What differentiates AI machines from other technology is the element of intelligence.

This allows rational agents to be able to make decisions while considering several variables in their environment. Machine learning is a subset of artificial intelligence in which a computer or AI model can learn from data and evolve on its own. Tom Mitchell defined machine learning in this way in 1997: “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.” Autonomous models/systems can to plan and decide sequences of actions without being told to do so. An autonomous machine does not necessarily have to have an element of machine learning in its functionality, nor do all AI systems use machine learning, which is a common misconception.

AI is leveraged in many different ways across industries and business sectors globally. Within healthcare, AI has been able to assist in several facets, such as medical diagnostics, imaging, drug discovery, and personalized medicine. Deep learning models, a type of machine learning model based on the brain that uses extensive layers of real-number representations, are commonly used to detect patterns in large medical image datasets. In finance, AI is used copiously to predict market trends and assist in stock trading decisions. Fraud detection is something that has also heavily relied on AI to increase efficiency and pattern recognition. (Weng et al., 2024) AI models can easily spot odd transactions or potential scams and phishing ploys. In addition to healthcare and finance, AI is prevalent in manufacturing, retail, software development, education, and more. The wide range of use cases that AI has implicates a larger population as one starts to think about the effects of AI regarding trust and agency.

There are considerable shortcomings when it comes to AI use despite all it can and has already achieved. Some of these include bias and trustworthiness. Bias in the context of AI refers to decisions made by AI systems that deviate from fairness or accuracy due to biases in their design or data. Data bias occurs when the data used to train a particular model is unrepresentative of the broader population. This type of bias manifests itself when errors are made in data collection or when the data used reflects historical or institutional biases that were held when the data was collected. Algorithmic Bias occurs when the design of an algorithm itself perpetuates certain biases. This can occur mainly when certain assumptions that do not hold across different contexts are used as the base of an algorithm. (Oyeniran et al., 2022) Regarding trustworthiness, accuracy has historically been the main factor in determining how trustworthy an AI model is, which can be misleading and neglect other components, such as ethics. To ensure the assessment of trust is more holistic, the AI community has identified these elements as additional criteria for trustworthy AI: robustness, security, transparency, fairness, and safety. (Li et al., 2023) These particular elements will be taken into consideration when discussing how to ensure AI systems are held to a high standard of trustworthiness in business environments, as well as being able to address other shortcoming such as bias.

STS Framework (ANT):

In "Where Are the Missing Masses," Bruno Latour discusses the impact that nonhuman "actors" have in several different systems. He uses door hinges, seatbelts, road signs, and keys to explain the role that nonhuman actors play in our everyday lives and interactions. These actors, though not sentient, affect the way people interact with their environments and perceive the world around them(*Where Are the Missing Masses, Sociology of a Few Mundane Artefacts | Bruno-Latour.Fr*, n.d.). This acknowledgment is the foundation of Bruno Latour's Actor-Network Theory (ANT). Actor-network theory is defined as a framework for understanding how both human and non-human entities interact to shape social and technological systems. The theory states that for an actor to act, many others must as well, implying that a collective effort drives action. (Bencherki, 2017) There is also an emphasis on the actor itself being a complex network of its own as well, further highlighting the effect that a nonhuman actor could potentially have on a particular system or network.

ANT is a very commonly referenced framework, which has also led to it receiving a myriad of critiques. This led even Bruno Latour, one of the creators of ANT, to publish an article addressing some of the common criticisms of the framework. Some of these criticisms include the broadness of ANT and its connection to social issues. Latour addressed concerns specifically on how networks were referred to and how ANT was utilized by many writers to describe social relations. Many would interpret the "networks" in ANT as technical networks that resemble technological interactions and systems. Latour goes on to say that the networks that he refers to in ANT are more complex and multi-faceted. He states, " A technical network in the engineer's sense is only one of the possible final and stabilized states of an actor-network. An actor-network may lack all the characteristics of a technical network - it may be local, it may have no compulsory paths, no strategically positioned nodes" (Latour, 1996). This illustrates much of the depth of connection that is possible with ANT. Latour also addresses the use of ANT to analyze social networks by clearly stating that ANT has little to do with the study of social networks but rather aims to describe the nature of societies by extending the role of an actor or actant to non-human entities. (Latour, 1996) There has also been criticism about the anti-dualism that is central to the ideas of ANT. Writers have claimed that the need to distinguish the separation between non-human and human actors to show their interconnectedness recreates the dualism that it is combating. (*Is Actor Network Theory Critique?*, n.d.)

Despite its critiques, ANT has been used as a framework to include many pieces of technology in critical analysis and frame more complex outlooks on society regarding the role technology plays in human interactive systems. An example that connects the ideas of ANT to the issue of trust and agency in AI is the Human-Machine-Network (HMN). A HMN is a networked system in which humans and machines interact, generating synergistic and often distinctive outcomes. Both humans and machines are considered actors, as they each engage with at least one other actor within the system. (Engen et al., 2016) In relation to AI, one can use HMNs to model most interactions with AI in personal or professional spaces. HMNs define an understanding of both machine and human agency by defining machine agency in relation to how creative the activity the machine is participating is, the influence the machine has on other actors, and the extent to which they are perceived as having agency by human actors. (Engen et al., 2016) In the context of AI, which has a plethora of different use cases, HMNs can help deeply analyze the role that AI systems can have, depending on the different tasks they are being leveraged to perform.

Analysis:

Several components can be focused on to address the issue of trust in AI-driven applications and businesses. These include explainability, ethical AI principles, and bias within AI usage. All of these factors should be addressed to build trust in AI use in business, and they all connect to specific actors identified that are part of this network of business functionality as a whole. These identified actors include a business entity, a specific employee within a business, and an AI model. In this network, the corporation acts as the primary initiator, deploying AI systems to enhance productivity and decision-making processes. Employees, as key stakeholders, play a crucial role in the acceptance and effective utilization of AI technologies, with their trust levels significantly impacting the success of AI implementation. The AI model serves as a non-human actor, influencing workflows, decision-making processes, and overall business operations. The interactions between these three actors form a dynamic network where trust acts as a binding force, determining the extent to which AI is embraced and integrated into the workplace. Understanding the relationships and translations between these actors is essential for addressing ethical concerns, mitigating risks, and fostering a culture of trust that enables the full potential of AI in business settings.

How all actors fit together:

Though building trust in AI affects the three actors mentioned above in different ways, all the actors are crucial to the proper functioning of the other. The corporation itself and the high-ranking executives and board members who contribute heavily to decision-making are essential for establishing the general vision and culture of a company. In regard to AI development, they can decide how much the company as a whole will incorporate AI into its work. The culture and priorities that the corporation sets trickle down to the employees. Middle management directly communicates specific expectations and goals that are heavily influenced by the culture of those above them. This will, in turn, affect the actual work that their subordinates are producing for the company and how they engage with artificial intelligence in the process. The model itself has to perform in such a way that the employees using it can not only produce serviceable work but also feel like they can trust and depend on AI models reliably. The data that the model is trained on affects the information given to the model, and its algorithms can inform the thought processes of employees if properly implemented and transparently communicated. Ultimately, the interplay between these three actors—corporate leadership, employees, and the AI model—creates a feedback loop where trust in AI is both shaped by and essential to the organization's overall success.

Transparency and Explainability:

Artificial Intelligence’s functionality separates it from many traditional technologies. The capacity for AI to engage in autonomous tasks and activities adds additional risk and uncertainty to many of its actions. In addition, the specific functionality of many AI models is extremely complex and hard to understand, even for more technical employees. This is what makes explainability extremely important to the increase of trust in AI applications. Explainability in this context is the ability of an AI model to explain its actions clearly to a user. By increasing a user's trust, the algorithmic recommendations are perceived as more useful and legitimate (Choung et al., 2023). Explainability would also increase the likelihood of a business entity adopting widespread use of AI within the business. Information regarding these technologies needs to be provided in a digestible manner for those without technical knowledge to be able to trust it with business processes. This would increase the *inductive* trust that executives have with AI, with *inductive* trust meaning trust from personal experience (either their own direct experience or observed experience from other entities or people). (Andras et al., 2018) These observed experiences are likely the only ones that these high-level executives or board members have had with AI, and having AI models with easily explainable functionality allows for trust to be granted from the top down. For a more hands-on and directly involved employee who is either in an entry-level or mid-level management position, interactions with AI are likely to be much more firsthand. For these employees, explainability is not enough, there is also a need for interpretability within a corporation. An AI system is interpretable if it can clearly produce information about how it reached a certain conclusion or developed a certain result. (Andras et al., 2018) Because employees are using this software directly, the interpretability of explainable AI is crucial for comfortable use, as well as effective collaboration with other workers. The ability for a worker to understand the purpose of the AI they are using is crucial to its effective use and provides a more solid basis for meaningful work within a team. (Continue with systems and entropy)

AI systems themselves frequently operate in a manner that is very much a “black box” (unclear to observers), which in turn can produce plenty of disorder and confusion (Mylrea & Robinson, 2023). Computer Scientists refer to disorder as something called *entropy*. Entropy is a quantifiable metric in the computing field and can usually be calculated with certain logarithmic functions. High amounts of entropy or disorder cause trust in a system to decrease because these systems are too random and complex, thus making them harder to explain. However, systems with entropy that is too low seem to be less intelligent and capable due to higher levels of predictability. (Mylrea & Robinson, 2023) Due to this, a balance between enough complexity that allows for intelligent output and enough simplicity for appropriate comprehension must be found. Allowing users to be able to tweak AI models to find what balance works for them may be a step in the right direction for users to have the most trust possible in the model.

Ethical issues (agency + security):

Ethical concerns regarding artificial intelligence are a dime a dozen. With how fast AI is developing and how much is still to be discovered about AI, there are many ethical considerations that need to be accounted for when using AI in businesses. Two major ethical issues concerning AI are machine agency and privacy. As non-human entities, the moral capacities of AI models are particularly obscure; however, due to their programmed functionality through machine learning algorithms and patterns, they have a level of agency and autonomy as well (Zhou et al., 2020). What adds further complexity is the access that AI can have over personal information and data. When using artificial intelligence in an application, one may train it on large datasets or use it on systems that contain massive amounts of sensitive information pertaining to a particular company, client, or government. Especially if an AI model does not have a certain level of explanatory power (as mentioned above), it can be hard to know what the reasoning behind the actions of an AI model is or what it plans to do with whatever information it becomes privy to. This can affect the autonomy and invade the privacy of whatever entity is in possession of said information. (Bankins, 2021) When thinking of how a business needs to take these issues into account, high-ranking executives need to consider how much control they want to maintain over the functionality of their business. They are already delegating work and management through robust organizational structures, but with AI in the picture, it works itself into this preset system. The difference comes in the humanity of who is receiving the delegated work, because when you delegate work to an AI system whose intentions are difficult to access, that in turn potentially takes away control and knowledge of company work from executives. These issues need to be addressed at an organizational level in order for the vision of an entity to remain robust. Many corporations have committees and even ethical codes to address this, but depending on how much AI is used at a company, a dedicated employee or even a department companywide dedicated to AI could ensure that AI is given the proper moral attention (Zhou et al., 2020). This additionally affects employees in a similar way. Accountability for tasks done using AI can become unclear to assign, regardless of whether its effect was positive or negative. This also affects the agency of employees over their own work that they produce. AI models can establish trust with human users by emphasizing their non-human identity. This may help make those using it more comfortable by reminding users that the agency of the AI machine they are using does not have any human thought processes attached to it. This is also essential to simply just remind users that it is a machine and should be perceived as such (Andras et al., 2018).

Talks about bias

Bias is another element of artificial intelligence that can heavily affect its accuracy and functionality. Additionally, bias in AI can have social implications that can perpetuate damaging ideologies in addition to producing incorrect information. This can spell bad news for a corporation that does not actively combat this, as whatever work they could produce could not only be lacking in quality but also alienate entire demographics of employees or clientele. Bias in AI opens an organization up to potentially damaging risks that could put financial strain on a corporation as well as possible legal trouble (Oyeniran et al., 2022). Organizations need to address bias at its source and ensure that all sectors of their business are devoted to eliminating bias from their work. On a higher level, top-level executives cannot necessarily do much to directly combat bias in AI systems, so it is imperative that they are aware of the damage it can have on their company and ensure that it is a priority for their mid-level management to monitor more hands-on employees. This can come in the form of mandating and participating in specific certifications or bias training to mitigate bias affecting data interpretation and the analysis of AI output.

For employees who work hands-on with AI systems on a frequent basis, explicit action can be taken. Three main sources of bias in AI can be addressed through direct action, and these include data, algorithmic, and human factors that introduce bias into the output of AI models. To address data-related bias, several techniques, such as using stratified sampling to guarantee all subgroups of a population are included in a dataset or actually analyzing data sets already in use to see if bias already exists in them (Oyeniran et al., 2022). Data-related bias mainly originates from the collection of data, so ensuring that if you are collecting data from third parties, they are prioritizing ethical methods of compilation is crucial as well. To combat algorithmic bias, fairness metrics can be included in the processes of model evaluation. This is particularly useful if, for some reason, a dataset that was shown to have bias is used. Human bias can be combated with regular audits and monitoring of AI systems. These audits would check for indications of unintended bias and attempt to find patterns that could be avoided in future work (Oyeniran et al., 2022).

Conclusion:

As corporate entities increasingly integrate AI-driven systems into their business operations, building and maintaining trust becomes a critical challenge. This trust is strengthened through three key areas: explainability, bias mitigation, and ethical considerations surrounding privacy and agency. By prioritizing explainability, organizations can ensure that AI models are transparent, making their decision-making processes more understandable and predictable for employees and consumers alike. Additionally, addressing bias mitigation is essential to creating fair and equitable AI systems, preventing discriminatory outcomes that could erode public confidence. Finally, corporate entities must uphold ethical principles of privacy and agency, ensuring that AI respects user autonomy and safeguards sensitive data. Ultimately, a commitment to these pillars will not only foster trust but also enhance AI's effectiveness, paving the way for more responsible and sustainable AI adoption in business.

Bibliography

Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., Urquhart, N., & Wells, S. (2018). Trusting Intelligent Machines: Deepening Trust Within Socio-Technical Systems. *IEEE Technology and Society Magazine*, *37*(4), 76–83. IEEE Technology and Society Magazine. https://doi.org/10.1109/MTS.2018.2876107

Bankins, S. (2021). The ethical use of artificial intelligence in human resource management: A decision-making framework. *Ethics and Information Technology*, *23*(4), 841–854. https://doi.org/10.1007/s10676-021-09619-6

Bencherki, N. (2017). *Actor–Network Theory* (C. R. Scott, J. R. Barker, T. Kuhn, J. Keyton, P. K. Turner, & L. K. Lewis, Eds.; 1st ed., pp. 1–13). Wiley. https://doi.org/10.1002/9781118955567.wbieoc002

Carew, S., Cooper, A., & Banerjee, A. (2025, January 28). DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss. *Reuters*. https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27/

Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction*, *39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Engen, V., Pickering, J. B., & Walland, P. (2016). Machine Agency in Human-Machine Networks; Impacts and Trust Implications. In M. Kurosu (Ed.), *Human-Computer Interaction. Novel User Experiences* (pp. 96–106). Springer International Publishing.

https://doi.org/10.1007/978-3-319-39513-5_9

Is Actor Network Theory Critique? (n.d.). https://doi.org/10.1177/0170840607082223

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.*, *55*(9), 177:1-177:46. https://doi.org/10.1145/3555803

Mylrea, M., & Robinson, N. (2023). Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI. *Entropy*, *25*(10), Article 10. https://doi.org/10.3390/e25101429

Oyeniran, O. C., Adewusi, A. O., Adeleke, A. G., Akwawa, L. A., & Azubuko, C. F. (2022). Ethical AI: Addressing bias in machine learning models and software applications. *Computer Science & IT Research Journal*, *3*(3), Article 3. https://doi.org/10.51594/csitrj.v3i3.1559

Weng, Y., Wu, J., Kelly, T., & Johnson, W. (2024). *Comprehensive Overview of Artificial Intelligence Applications in Modern Industries*. https://doi.org/10.48550/ARXIV.2409.13059

What is DeepSeek—And why is everyone talking about it? (2025, February 4). https://www.bbc.com/news/articles/c5yv5976z9po

Where are the missing masses, sociology of a few mundane artefacts | bruno-latour.fr. (n.d.). Retrieved February 17, 2025, from http://www.bruno-latour.fr/node/258.html

Zhou, J., Chen, F., Berry, A., Reed, M., Zhang, S., & Savage, S. (2020). A Survey on Ethical Principles of AI and Implementations. *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 3010–3017. https://doi.org/10.1109/SSCI47803.2020.9308437