Do We Trust AI: An Investigation into Public Sentiment about Artificial Intelligence and its Trustworthiness

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 of the Degree
Bachelor of Science, School of Engineering

Kevin Ming Chung

Spring 2024

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

Kent Wayland, Department of Engineering and Society

Introduction

Imagine you are in the passenger seat of a driverless autonomous vehicle. Imagine you show up to the doctor's office, only to be diagnosed by a robot using artificial intelligence (AI). Are you comfortable in either of these situations? Do you trust the AI working behind the scenes? Most people would probably answer "no" to both of these questions. According to the annual American Automobile Association's autonomous vehicle survey, 91% of Americans stated that they do not fully trust autonomous vehicles (Moye, 2023). There have been numerous other studies that have shown similar results of people not trusting AI (Gillepsie et al., 2023). Without trust, we may not be able to take full advantage of the benefits that AI can provide.

Trust has been shown to be essential for a technology to be adopted (Hoff and Bashir, 2015). AI is extremely powerful, and has already improved performance in fields such as chess, medical diagnoses, and more (Kaur et al., 2020). However, due to the fact that AI is still a developing technology, it is understandable that many hesitate to trust it. By building trust in AI, it might allow AI to flourish in many more industries. In order to do so, we need to understand the reasons for the lack of trust in AI. In this paper, I will investigate public sentiment on AI with regards to its trustworthiness, and the reasons behind it. When referring to the public, I am referring to anyone who is concerned with the use of AI, regardless of expertise in the subject.

Definitions

IBM defines AI as "technology that enables computers and machines to simulate human intelligence and problem-solving capabilities" (n.d.). Cutting-edge AI is the closest we have to human intelligence for computers, but there is still a large difference between human and artificial intelligence. AI tends to excel at problems that can be quantified, and is very specialized for a certain task, such as detecting pedestrians on the street. AI models also require

much more examples to learn something than a human. On the other hand, humans are able to think critically and reason about a situation, where AI cannot. One day, AI may be able to completely replicate human intelligence, but it still has a long way to go.

AI is a very broad term that encompasses many different things. While it may be useful to more specifically define AI in most cases, it actually makes more sense to leave this definition broad. My research is not concerned with the more technical definition of AI, as when people discuss AI online, they are often using the more general, colloquial understanding of what AI is. Limiting the definition to a more specific and technical one would exclude the opinion of those who are less knowledgeable about the specifics of AI, and these opinions are just as important as those educated about AI, since they could also be users.

As for trust, exploring the definition of trust and the various forms of trust is outside the scope of this paper. For my research, I will be using an established definition of trust laid out by Jacovi et al., who formalized trust in artificial intelligence (2021). Derived from research on interpersonal trust, they developed two key elements to human-AI trust: anticipation and vulnerability. Anticipation refers to a belief that AI will act in the human's best interest. This also means that in order for trust to be verified, there must be a risk of an undesirable or disadvantageous outcome to the human in a given situation. In this case, acting in the human's best interest is to avoid any undesirable outcomes. Vulnerability refers to the human willingness to be subjected to the AI's actions. When humans trust technology, there is an inherent vulnerability, as users are allowing technology to act on their behalf. When there is no vulnerability between humans and AI, there is also a lack of trust because the users do not necessarily believe that the AI will act in their best interest. It's important to note that this does

not mean that there is distrust, as people can have a more neutral stance on the trustworthiness of AI.

Sociotechnical Context

AI has already worked its way into many fields and has provided lots of benefits. For example, there is research showing that AI can provide a positive impact on medical diagnoses (Kaur et al., 2020). However, despite all of its promise and growth, there is a lack of public trust in AI. According to a survey of over 17,000 people across 71 countries conducted at the University of Queensland, three out of every 5 participants were hesitant to trust AI systems. Some reasons cited were doubts about security, fairness, and the safety of AI systems (Gillespie et al., 2023).

Hoff and Bashir, in a study analyzing factors influencing trust in automation technologies, wrote that trust is one of the most important factors when determining if a technology is going to be adopted (2015). Often when adopting a new technology, the potential adopter does not have complete information or control over the technology, and a "leap of faith" must be made in adopting that technology. That "leap of faith" requires some sort of trust in that technology (Bahmanziari et al., 2003). Without trust, we may not realize all of the benefits of AI. Determining the reasons behind the current public lack of trust in AI will help researchers and developers address those concerns and ensure that AI in the future will provide the promising benefits that it shows without harm.

Ensuring AI acts in our best interest and is able to take into relevant ethical considerations in any given situation is essential to the safety and security of both the AI and those that interact with it (Winfield and Jirotka, 2018). If, for example, there was an AI-based medical tool that diagnosed patients, we would want to feel safe disclosing any sensitive

personal information to it. Terrasse et al. showed that moving towards virtual medical interactions help some patients feel more comfortable sharing sensitive information, which would indicate that some may trust a nonhuman with such information. However, these interactions tend to depersonalize patient-practitioner interactions, which have a therapeutic effect that can be beneficial to the patient (2019).

There has been a lot of research into the antecedents of trust in technology. There is a consensus that the most important factor that influences trust in technology is the reliability of a technology (Jacovi et al., 2021; Lockey et al., 2021; Sutrop, 2019). Due to the fact that AI is still a developing technology, it is difficult for those who do not work with AI to truly evaluate the reliability of AI models (that are not already widespread). The ability to do so is essential to trust, as trust for no good reason is not trust at all, but rather blind faith (Lockey et al., 2021). One of the reasons most people cannot evaluate AI properly is because of the "black-box" nature of some AI models (Rossi, 2018). A solution to this is to require AI to be certified and approved before use, as many other technologies are (Sutrop, 2019).

Methods

In order to analyze public sentiment about trust and AI, I have collected videos, articles, and internet posts that share public thoughts about AI and its trustworthiness. The main sources for this data are from websites such as Youtube, Reddit, and online articles. In this case, "online article" refers to a short essay found on a blog, newsletter, or news website. Keeping this search as rigorous as possible, I conducted the search on a private browser, and used many search prompts with positive, negative, and neutral sentiments towards the trustworthiness of AI. Some examples of searches are "Why should we trust AI", "Why shouldn't we trust AI", and "Is AI trustworthy?". Sources were filtered out if they do not provide reasons for their stance on

trustworthiness and AI or if they do not address the topic of trustworthiness and AI at all. They were also filtered out if they are more than two years old, as the landscape of AI is rapidly changing. In order to select which sources to use for data, I relied on Google to provide relevant and prominent results – I tended towards the earlier results in each search. This introduces more bias towards larger websites with more views, but these websites often help shape public opinion due to the exposure, and still serve as a decent idea of public sentiment.

Once all the sources were collected, I distilled them down to whether they have positive or negative sentiment towards AI, and the reasons for believing so. This is done through inductive coding with codes centered around reasons for trust or lack of trust. Coding is a process in qualitative studies that involves organizing data by themes and relationships. Inductive coding means that the codes are created after the data has been found; I deemed this more useful than deductive coding – where codes are created before data collection – since I was unsure of the reasons behind trust or lack of trust in AI.

After the sources were distilled and coded, I synthesized them into a collection of reasons for trust or lack of trust in AI. For each reason found, I determined whether they showed anticipation and/or vulnerability. I also compared them to the reasons found in previous research for a lack of trust in AI. By doing so, it gives an understanding of the general sentiment towards AI in terms of trust and the leading reasons behind that sentiment.

It is important to note that these methods are not ideal. By collecting data from articles, Youtube videos, and social media posts, there is a strong bias towards content from creators, influencers, and journalists who have a desire to publicize their opinions. Additionally, it is nearly impossible to capture a stratified sample of opinions on the internet, since search engines are more likely to give results that have been viewed by many people already. This clearly

cannot completely or truly capture the state of the public sentiment on AI and the reasons for it, but in the scope of this study, it likely provides a good idea of the common reasons and themes present in the public sentiment about AI. Ideally, I would be able to conduct a rigorous survey and distribute it to a stratified sample of participants, but that was unfeasible with the time frame and resources available for this study. The intent of this research is not to provide a complete scientific analysis on public sentiment with regards to trust and AI, but rather to take the temperature of that sentiment and provide direction for future research.

Results

For my analysis, I collected 10 Youtube videos, 10 articles, and 10 Reddit posts that all met the criteria outlined previously. Each of the Youtube videos was between 2-20 minutes long and widely varied in popularity and depth of explanation (TEDx Talks, 2023; TEDx Talks, 2024; English with Cambridge, 2023; The Daily Aus, 2022; Dr Waku, 2024; IBM Technology, 2024; Rise of AI, 2023; University of Oregon, 2023; Modern AI, 2023; Institut for Datalogi, Aalborg Universitet, 2023). The Reddit posts were all comments, as most of the opinions on Reddit are shared in response to an initial question or hypothetical (miss3lle, 2023; whyisitsooohard, 2024; vVveevVv, 2022; Tom_Bombadil_1, 2024; ayleidanthropologist, 2022; hvgotcodes, 2023; Mass_Emu_Casualties, 2022; QuicksandHUM, 2023; incoherent1, 2024; [Reddit comment about trusting AI in healthcare], 2022). The articles came from various blogs and new sources (Hanson, 2023; Pearce, 2023; Thurston, 2023; Schneier & Sanders, 2023; Bailey, 2023; Senko, 2023; “Is it possible to trust Artificial Intelligence (AI)?”, 2022; Leffer, 2024, “Why scientists trust AI too much – and what to do about it”, 2024; Li, 2023).

It is pretty much impossible to sort the data into two categories of “trusts AI” and “doesn’t trust AI”. None of the Youtube videos or articles were completely for or against trusting

AI. All of them approached AI with a level of skepticism, but believed that AI could bring lots of benefits to the world. There were, however, 3-5 Reddit posts that had negative sentiment towards AI, two with positive sentiment, and 3-5 with a neutral sentiment. I put ranges on the negative and neutral sentiment counts because it is not clear which category some of the posts fall into.

There were many different reasons outlined in each article, video, or post for the stance that the author took. Most of these were reasons for lack of trust in AI, but some were reasons for trusting AI. I organized the reasons into categories and provided counts of how much each reason appeared in a video, article, or post.

Table 1

List of reasons for lack of trust in AI and the number of times they appear in the data

Reason for lack of trust	**Youtube**	**Articles**	**Reddit**	**Total**
Bias/Prejudice	5	4	2	**11**
“Hallucinations”	3	4	2	**9**
Potential malicious use	2	1	3	**6**
Lack of public awareness of limitations of AI	3	2	1	**6**
“Black box”/Lack of explainability	2	4	0	**6**
Lack of accountability	4	1	0	**5**
Lack of transparency about model construction process	2	2	0	**4**
Data privacy	3	0	0	**3**
Media portrayal of AI	0	0	2	**2**
Lack of morality/ethical understanding	1	1	0	**2**
AI only understands correlation, not causation	1	0	0	**1**
Uniqueness neglect	0	1	0	**1**

Table 2

List of reasons for trust in AI and the number of times they appear in the data

Reason for trust	Youtube	Articles	Reddit	Total
Belief of objectivity	1	2	1	**4**
Necessity due to integration	1	1	0	**2**
Ability to account for more factors than humans	1	0	1	**2**

Analysis/Discussion

From the results, it is abundantly clear that there is more skepticism than trust in AI from opinions on the internet. This was true even when I searched with prompts that had positive sentiment in AI, such as "why can we trust AI?" – almost all of the results were still reasons for not trusting AI. However, none of the sources called for an abandonment of AI or development in AI. The consensus was that AI has the potential to be extremely beneficial to our society, but should be approached with caution, and should be seen as a tool or an augmentation rather than a replacement for human decision-making.

Let's first examine the most common reason for lack of trust in AI – bias/prejudice. One of the reasons for this is because AI models are built to understand extremely complex correlations. Any bias or prejudice that exists in the data used to construct a model, even if it is completely unintentional, could potentially be propagated in the models' output. In an article by Dr. David Leslie of the Alan Turing Institute, he lists several examples of AI-based facial recognition algorithms that perform better on white faces than black faces. This is due to the training datasets for these models having a significantly larger number of white faces than black

faces (2020). Developers are responsible for the software they create, and many of the articles in the data called for more accountability for the developers when bias exists in AI.

The second most popular reason for lack of trust in AI can generate false information, or "hallucinations", as it is known in the field of generative AI. This point mainly pertains to just generative AI, since large language models (LLMs) such as ChatGPT are seemingly capable of providing whatever information you ask of it. The way that ChatGPT is trained, however, does not enforce that the information provided is realistic, only that it appears realistic. There is an example of a lawyer using ChatGPT for a legal brief in court, and ChatGPT fabricated names of cases that seemed realistic (Maruf, 2023). Generative AI for images is entirely "hallucinations", as it is designed to create realistic-looking images or art that aren't actually real or created by a human.

Both of these two reasons clearly show a lack of trust, as anticipation is not present and vulnerability is dangerous. As a reminder, anticipation refers to the belief that technology will act in the best interest of the user, and vulnerability is allowing that technology to do so. With bias or prejudice present in models, it cannot act in the best interest of all users, thus lacking anticipation as it pertains to trust. While there may not be much risk in every situation where AI is used, something as important as a medical diagnosis could lead to severe consequences if there is bias. In that context, the patient is vulnerable to the decision of the AI, and that vulnerability could be betrayed if the AI makes the wrong decision. "Hallucinations" also result in a lack of the two aspects of trust. Assuming a user that is seeking accurate information, an AI that spreads incorrect information is not acting in the interest of the user, meaning that the user's vulnerability could be betrayed.

Most of the AI models used today are “black boxes”, another one of the common reasons cited for lack of trust in AI. A “black box” in this context means that we can control the input of a model and see its output, but we have no idea of how it gets to that output. With human-to-human trust, even if we don’t necessarily agree with someone’s decision, we can ask them to explain their reasoning and maintain trust. Because we cannot do the same with AI, we cannot trust it the same that we do with other people. The field of explainable AI (XAI) is dedicated to this, where researchers are trying to create more transparent models or models that can explain another model’s decisions (Xu et al., 2019). XAI is quite new in AI research, and still has a long way to go before it can explain the output of the large models being used today. Associate professor Andres Masegosa of Aalborg University believes that there is a tradeoff between accuracy and explainability in the current state of AI. We are able to explain much simpler models, but they cannot capture the complex trends that a more complex, less explainable model can (Institut for Datalogi, Aalborg Universitet, 2023).

Some of the less common reasons listed in the data for a lack of trust in AI are very interesting. Media portrayal of AI in news, TV, and movies almost always poses the hypothetical situation of AI taking over the world. While this may not be a threat currently, it strikes fear into the public and sets expectations for AI that are unrealistic for the technology of today. Another reason for lack of trust, uniqueness neglect, really piqued my interest. Uniqueness neglect refers to the belief that some people have that their situation is “too unique” for AI to capture, and that only a human could understand what to do with their situation (Thurston, 2023). This is most prominent in medical contexts, where people want a human doctor to analyze their situation because they believe it is unique.

While the overwhelming majority of the reasons were for a lack of trust in AI, there were still a few for trusting AI. Many people believe that AI is objective or lacks bias. The datasets that it is trained on will always have bias, and given the current way that models are constructed, they will almost always have the same bias present in the dataset. Believing AI is objective can lead to initial trust, but this kind of trust is bordering the line of "blind faith".

One of the other reasons that people tend to trust AI more than humans in some scenarios is the ability for AI to factor in almost every aspect of a situation. By accounting for so many factors, AI can sometimes find trends and patterns that are essentially invisible to humans. While this holistic ability may contribute to trust, it still does not seem to outweigh the factors that lead to a lack of trust. AI can take in many factors, but we still have no idea how it is using those factors or which factors are the most relevant to the decision at hand.

The last reason that I found in the data for trusting AI is probably the most interesting. AI is being integrated more and more into our daily lives and will continue to be used for more. Mark Bailey, chair of cyber intelligence and data science of the National Intelligence University, believes that at some point in integration, human intervention is impossible or extremely difficult, and there is no choice but to trust the AI in order to use a product or service (2023). For example, we have to trust that the AI that handles facial recognition for our phones will keep our phones secure, since it is the only choice that phone manufacturers provide (besides of course, the passcode). Bailey's argument is that if we get to the point where AI is controlling our infrastructure, then we will need to trust that AI in order to use things like the internet or electricity. Bailey's definition of trust in this article is different from the one outlined in this paper. I would call this *reliance* rather than trust, since I have defined trust as believing that something will act in your best interest. Being reliant on something doesn't necessarily mean that

you trust it, but the situation that Bailey poses would make it very difficult to act upon one's lack of trust in AI.

Comparing the data to reasons found in previous research, a lot of the same reasons are present. The survey conducted by Gillespie et al. found that doubting AI's fairness was one of the main reasons for the lack of trust (2023). This is very similar to the most common reason found in my data, as a model with bias or prejudice is likely not fair. They also cited doubts in security and safety as other reasons, which was somewhat present, given the couple occurrences of the data privacy reason in the data. However, the survey was more focused on using AI in the workplace, where data security is often a higher priority. The "black-box" nature of AI cited by Rossi was also quite present in the data (2018). The similarities in the data collected not only helps confirm the findings of previous research, but strengthens the generalizability of future research.

Conclusion

The public sentiment around AI with regards to trust is one of skepticism. AI is still a very young technology, and we are experiencing the growing pains as we develop it more and use it in more ways. By taking the time to identify and understand the reasons behind the current skepticism, we can help developers improve AI to be more suited to what the public wants. Transparency, privacy, fairness, and accountability are some of the important values that people have pointed out and advocate for in AI. Continuing this advocacy is also essential for the development of AI, and users should also hold developers accountable for violations of those tenets or for misuse of AI.

This paper has only scratched the surface of understanding how people feel about AI; future research could investigate this topic much more in-depth by running experiments or

widespread surveys and interviews. There are also likely lots of people who do feel like they trust AI, but do not share their opinions online, and a future investigation should definitely explore that a lot more to determine what is good about AI. It's a very nuanced topic, and we should ensure that a discussion continues between developers and users of AI, so that it's able to benefit everyone.

References

ayleidanthropologist. (2022, August 1). *One day yes, 100%. But right now the AI generated stories and headlines are just too fresh in my mind.* [Comment on the online forum post *Would you trust a robot to examine, diagnose and prescribe a treatment for you? Would you trust a robot as your physician or surgeon, if your doctor or family or friends suggests? – A Study on Human – AI trust.*]. Reddit. www.reddit.com/r/compsci/comments/wdiuon/would_you_trust_a_robot_to_examine_diagnose_and/iiin4xw/.

Bahmanziari, T., Pearson, J. M., & Crosby, L. (2003). Is trust important in technology adoption? A policy capturing approach. *The Journal of Computer Information Systems*, *43*(4), 46–54.

Bailey, M. (2023, October 3). *How Can We Trust AI If We Don't Know How It Works*. Scientific American. https://www.scientificamerican.com/article/how-can-we-trust-ai-if-we-dont-know-how-it-works/

Dr Waku. (2024, January 7). *Can We Trust Decisions Made by AI?* [Video]. YouTube. https://www.youtube.com/watch?v=ODHxhNTfqEg.

English with Cambridge. (2023, April 3). *Can We Trust AI?* [Video]. YouTube. https://www.youtube.com/watch?v=yL7x-GqeF78.

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Ali Akbari. (2023). *Trust in Artificial Intelligence: A global study*. The University of Queensland; KPMG Australia. https://doi.org/10.14264/00d3c94

Hanson, R. (2023, August 10). Can we trust A.I. to tell the truth? *Reason.Com*. https://reason.com/2023/08/10/can-we-trust-a-i-to-tell-the-truth/

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

hvgotcodes. (2023, May 3) *People saw The Terminator and think the natural endgame for AI is the machines try to kill us and take* [Comment on the online forum post *Can you guys please explain what are the genuine 'Dangers of AI'?*]. Reddit. www.reddit.com/r/AskScienceDiscussion/comments/136in7a/can_you_guys_please_explain_what_are_the_genuine/jiot5vl/.

IBM. (n.d.). *What is Artificial Intelligence (AI)? | IBM*. Retrieved March 28, 2024, from https://www.ibm.com/topics/artificial-intelligence

IBM Technology. (2024, February 19). *How to Build AI Systems You Can Trust* [Video]. YouTube. https://www.youtube.com/watch?v=mfxgfU5Abdk.

incoherent1. (2024, Feburary 28). *There is no possible way to avoid bias when LLMs are trained on information provided by humans who themselves are* [Comment on the online forum post *How can we trust AI or even an AGI if corporations incorporate their bias?*]. Reddit. www.reddit.com/r/singularity/comments/1b1wyv1/how_can_we_trust_ai_or_even_an_agi_if/ksi8pbw/.

Institut for Datalogi, Aalborg Universitet. (2023, August 11). *Why We Need Trustworthy AI* [Video]. YouTube. https://www.youtube.com/watch?v=BwPo5K8eKaw.

Is it possible to trust Artificial Intelligence (AI)? (2022, September 22). *Justice Everywhere*. https://justice-everywhere.org/technology/is-it-possible-to-trust-artificial-intelligence-ai/

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923

Kaur, S., Singla, J., Nkenyereye, L., Jha, S., Prashar, D., Joshi, G. P., El-Sappagh, S., Islam, Md. S., & Islam, S. M. R. (2020). Medical Diagnostic Systems Using Artificial Intelligence (AI) Algorithms: Principles and Perspectives. *IEEE Access*, *8*, 228049–228069. https://doi.org/10.1109/ACCESS.2020.3042273

Leffer, L. (2024, March 18). *Too Much Trust in AI Poses Unexpected Threats to the Scientific Process*. Scientific American. https://www.scientificamerican.com/article/trust-ai-science-risks/

Leslie, D. (2020). *Understanding bias in facial recognition technologies*. https://doi.org/10.5281/zenodo.4050457

Li, C. (2023, March 24). *Why We Don't Trust AI — and How to Change That | LinkedIn*. LinkedIn. https://www.linkedin.com/pulse/why-we-dont-trust-ai-how-change-charlene-li/

Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021). A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. *Hawaii International Conference on System Sciences 2021 (HICSS-54)*. https://aisel.aisnet.org/hicss-54/os/trust/2

Maruf, R. (2023, May 27). *Lawyer apologizes for fake court citations from ChatGPT | CNN Business*. CNN. https://www.cnn.com/2023/05/27/business/chat-gpt-avianca-mata-lawyers/index.html

Mass_Emu_Casualties. (2022, August 2). *Sure. Because as a women, male doctors tend to dismiss most of our symptoms or misdiagnose our problems because we* [Comment on the online forum post *Would you trust a robot to examine, diagnose and prescribe a treatment for you? Would you trust a robot as your physician or surgeon, if your doctor or family or friends suggests? – A Study on Human – AI trust.*]. Reddit. www.reddit.com/r/compsci/comments/wdiuon/would_you_trust_a_robot_to_examine_diagnose_and/iilusun/.

miss3lle. (2023, May 3). *Artificial intelligence can also be dangerous when used for decision making processes because it's aiming for a performance goal without* [Comment on the online forum post *Can you guys please explain what are the genuine 'Dangers of AI'?*]. Reddit. www.reddit.com/r/AskScienceDiscussion/comments/136in7a/can_you_guys_please_explain_what_are_the_genuine/jiptxki/.

Modern AI. (2023, September 26). *Why Should We Trust AI? Exploring the Ethics and Challenges #ai* [Video]. YouTube. https://www.youtube.com/watch?v=Snkem0kTDaw.

Moye, B. (2023, March 2). *AAA: Fear of Self-Driving Cars on the Rise*. AAA Newsroom. https://newsroom.aaa.com/2023/03/aaa-fear-of-self-driving-cars-on-the-rise/

Pearce, K. (2023, March 6). *Can we trust AI?* The Hub. https://hub.jhu.edu/2023/03/06/artificial-intelligence-rama-chellappa-qa/

QuicksandHUM. (2023, May 3). *The problem is making AI align its goals and value to what you want. Whoever creates the AI deeply influences* [Comment on the online forum post *Can you guys please explain what are the genuine 'Dangers of AI'?*]. Reddit. www.reddit.com/r/AskScienceDiscussion/comments/136in7a/can_you_guys_please_explain_what_are_the_genuine/jiqw0be/.

Rise of AI. (2023, May 14). *Prof. Dr. Joanna Bryson | No One Should Trust AI | Rise of AI Conference 2023* [Video]. YouTube. https://www.youtube.com/watch?v=YgZD74jIMJY.

Rossi, F. (2018). Building Trust in Artificial Intelligence. *Journal of International Affairs*, *72*(1), 127–134.

Schneier, B., & Sanders, N. (2023, July 20). *Can you trust AI? Here's why you shouldn't.* The Conversation. http://theconversation.com/can-you-trust-ai-heres-why-you-shouldnt-209283

Senko, L. (2023, November 22). *How To Trust AI*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2023/11/22/how-to-trust-ai/?sh=46ab953b2789

Sutrop, M. (2019). SHOULD WE TRUST ARTIFICIAL INTELLIGENCE? *Trames. Journal of the Humanities and Social Sciences*, *23*(4), 499. https://doi.org/10.3176/tr.2019.4.07

TEDx Talks. (2023, January 3). *Artificial Intelligence and Trust | Marcel Isbert | TEDxTUDarmstadt* [Video]. YouTube. https://www.youtube.com/watch?v=kw__3DCIJ8E.

TEDx Talks. (2024, February 2). I*n AI We Trust. But Should We? | Aaron Hunter | TEDxAbbotsford* [Video]. YouTube. https://www.youtube.com/watch?v=Z0gPI1zzq5Y.

Terrasse, M., Gorin, M., & Sisti, D. (2019). Social Media, E-Health, and Medical Ethics. *Hastings Center Report*, *49*(1), 24–33. https://doi.org/10.1002/hast.975

The Daily Aus. (2022, July 10). *Can We Trust Artificial Intelligence? | The Daily Aus* [Video]. YouTube. https://www.youtube.com/watch?v=_N0WwOGjhCs.

Thurston, A. (2023, February 8). *Can We Trust ChatGPT and Artificial Intelligence to Do Humans' Work?* Boston University. https://www.bu.edu/articles/2023/can-we-trust-chatgpt-and-artificial-intelligence/

Tom_Bombadil_1. (2024, February 8). *"oh wow, these scrolls seem to be a recipe for Mary Berry's Classic Special Scones! Amazing. These ancients were so* [Comment on the online forum post *Can we really trust AI to tell us things we can't verify ourselves?*]. Reddit. www.reddit.com/r/PhilosophyofScience/comments/1alw0o5/can_we_really_trust_ai_to_tell_us_things_we_cant/kphgfqo/.

University of Oregon. (2023, April 24). *The Ethics of AI Explained | Can We Trust AI?* [Video]. YouTube. https://www.youtube.com/watch?v=oPA7xMrz4dM.

vVveevVv. (2022, August 1). *I think that augmentation would be more favourable than complete replacement. But that's just speculation, and somewhat of a subjective* [Comment on the online forum post *Would you trust a robot to examine, diagnose and prescribe a treatment for you? Would you trust a robot as your physician or surgeon, if your doctor or family or friends suggests? – A Study on Human – AI trust.*]. Reddit.

www.reddit.com/r/compsci/comments/wdiuon/would_you_trust_a_robot_to_examine_diagnose_and/iija8ar/.

Why scientists trust AI too much—And what to do about it. (2024). *Nature*, *627*(8003), 243–243. https://doi.org/10.1038/d41586-024-00639-y

whyisitsooohard. (2024, February 28). *Everyone is discussi…* [Reddit Comment]. R/Singularity. www.reddit.com/r/singularity/comments/1b1wyv1/how_can_we_trust_ai_or_even_an_agi_if/kskgzsu/

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180085. https://doi.org/10.1098/rsta.2018.0085

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). *Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges* (pp. 563–574). https://doi.org/10.1007/978-3-030-32236-6_51

[Reddit comment about trusting AI in healthcare]. (2022, August 1). *Yeah robot could comb through my symptoms and available treatments easier than a doctor who can't remember everything* [Comment on the online forum post *Would you trust a robot to examine, diagnose and prescribe a treatment for you? Would you trust a robot as your physician or surgeon, if your doctor or family or friends suggests? – A Study on Human – AI trust.*]. Reddit. www.reddit.com/r/compsci/comments/wdiuon/would_you_trust_a_robot_to_examine_diagnose_and/iiiin6d/.