

Rational Reasons for Realistic Reliance (R^2R^2)

Victoria Tan

OSTEM Internship Program, NASA's Langley Research Center

This document seeks to illuminate the dangers of using the term **TRUST** in relation to artificial intelligence (AI). The concept of **RELIANCE** will be introduced as a suitable replacement. Several informal thought experiments are used to distinguish **TRUST** and **RELIANCE** from one another, and to propose possible use cases for the latter. The goals of the paper are to address concerns about widely-adopted paradigms surrounding AI, explain the dangers posed by said paradigms, and introduce methods to mitigate potential harm.

I Introduction

To paraphrase an excellent paper, anthropomorphism of AI¹ is a phenomenon produced when a system appears to possess human-esque qualities, stemming from the language used in reference to the system [1]. People are largely unconscious that this process occurs, or that their linguistic or behavioral cues are perpetuating anthropomorphism, leading to a range of possible pernicious effects [2, 3]. However, discussing the general issue of anthropomorphism is beyond the scope of this paper. Instead, this document will focus on one specific influence of anthropomorphism: the usage of the term **TRUST**² in relation to AI.

TRUST is a thorny topic. “**TRUST** in AI” is, necessarily, more so. There exist many philosophical approaches to the concept of **TRUST**, and detailed coverage of even a few of the most consequential is also not the focus of this paper. However, it will be beneficial to cover some more widespread and pervading definitions and theories in this document. An influential paper by Lee and See defines **TRUST** as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.” [4]. Other definitions of **TRUST**, such as the one set forth by Baier as a first approximation, rely on a will-based theory [5]. Recent literature further splits **TRUST** into different paradigms—normative, affective, and rational—arguing that for an artefact to truly be capable of being trusted, it cannot only satisfy requirements of a rational account of **TRUST** [6].

Beyond these various opinions on the definition of **TRUST**, there also exist many different attitudes towards the concept of “**TRUST** in AI” [7, 8, 9, 10, 11]. Some believe that AI possesses an unusual amount of independent decision-making and thus should be treated as an artefact capable of being **TRUSTED** and being **TRUSTWORTHY** [12]. On the other hand, many philosophers on **TRUST** believe that **TRUST** = **RELIANCE** + F, where F is some extra factor associated with an agent that possesses a will [13]. Regardless of what this extra factor is, AI

¹The common initialism AI will be used throughout this document because it has come to be widely accepted across various fields. My mentor is not fond of it because he is not fond of the term “artificial intelligence”, preferring “probabilistic automation” or “probabilistic automation systems” depending on context.

²Throughout the paper, words such as **TRUST**, **RELY**, and **RELIANCE** will be typeset in SMALL CAPS to emphasize attention to these terms of art.

in its current form neither has a will nor is an agent. Therefore, we cannot say it is possible to truly TRUST AI.

Even if this theory seems unappealing, it is not necessary to believe in it to see that the main problem with the discourse on TRUST is that a majority view does not exist [6, 7, 14, 15]. Due to the broad range of meanings that could be implied when the term TRUST is used in correlation to AI, it is entirely possible that two people, saying they “TRUST ChatGPT”, are thinking about and implying vastly different things. However, given the many normative implications and connotations that TRUST already carries, as well as the continuing debate about its definition, standardizing one meaning of TRUST as it relates to AI poses complications. Instead, it would be better to choose an entirely new and agreed-upon way of referring to what we call TRUST. In a technological and complex field, it is necessary to ensure vocabulary and ways of thinking are similar, to avoid misunderstandings. Therefore, this new term should be comparatively freer of value-laden propositions, sufficiently precise, and flexible when addressing our designated concept.

This paper will introduce the paradigm of RELIANCE, which may serve as a suitable replacement for TRUST, and proposes several informal thought experiments to determine how it may be used and challenged. Section II explores historical and current definitions of TRUST and the ways in which this term is used in relation to AI. Section III introduces the definition of RELIANCE. Here, we address important differences between TRUST and RELIANCE and explain the reasoning behind the greater suitability of the latter for usage for AI. Section IV addresses the necessity of having a safe system, and whether assured safety can justify our RELIANCE on it. Section V addresses potential concerns with the paradigms we set forth in this paper, and offers warnings about the manner in which we think about AIs. The paper concludes by discussing the necessity of changing the TRUST paradigm in relation to AI to that of a RELIANCE paradigm, and potential benefits it could set forth when used in relation to safety and doubt.

II Misuse of TRUST

The *Oxford English Dictionary* defines TRUST (the noun) as “Firm belief in the reliability, truth, or ability of someone or something; confidence or faith in a person or thing, or in an attribute of a person or thing.” The OED goes on to define TRUST (the verb) as “To have faith or confidence in a person, quality, or thing; to rely on.” Such definitions are suitable for colloquial use and what can be termed “usage shortcuts.” However, these everyday definitions of TRUST differ from more formal definitions, particularly how philosophers of TRUST define it [7]. The Stanford Encyclopedia of Philosophy explains that “For most philosophers, trust is a kind of reliance. . . plus some extra factor” [16]. Our concern is not figuring out what this factor is (as it is still highly debated) but rather knowing that such a factor exists which sets TRUST apart from RELIANCE.

Ryan’s paper proposes characteristics of TRUST which may be helpful in elucidating normative qualities that distinguish TRUST from RELIANCE [6]. These characteristics are described in the paper as the following:

- i. *A has confidence in B to do X.*
- ii. *A believes B is competent to do X.*
- iii. *A is vulnerable to the actions of B.*

- iv. *If B does not do X then A may feel betrayed.*
- v. *A thinks that B will do X, motivated by one of the following reasons:*
 - (a) *Their motivation does not matter (rational trust)*
 - (b) *B’s actions are based on a goodwill towards A (affective trust)*
 - (c) *B has a normative commitment to the relationship with A (normative trust)*

Before further explaining the importance of these qualities, it may be helpful to provide demonstrations of how TRUST is used to refer to varying subjects. The below examples are ordered in terms of our strength of belief in the sensibility of the sentences.

I trust my mother.
I trust my friends.
I trust my dog.
I trust my pet rabbit.
I trust my car.
I trust my computer.
I trust my Wi-Fi connection.
I trust my clock.
I trust the microwave.
I trust my screwdriver.
I trust my fork.
I trust my notebook.
I trust my pet rock.

Though it is not necessary to enumerate on the differences between every single sentence, it will be beneficial to explain why the sentence at the very top is much more sensible than the sentence at the very bottom. Let us extend these sentences to situate them within helpful contexts:

Example: I trust my mother to take care of my roses while I am on vacation.

Example: I trust my pet rock to keep steady so that I don’t fall.

If I return from a vacation to find out my mother has let my roses die, I might feel angry or upset. Hence, it would not be out of order for me to ask for an apology from her, express my emotions, or expect that she understands how I feel. However, if I climb on top of my pet rock and it wobbles, leading to me falling off and skinning my knee, I might be upset, but it would be nonsensical to berate the rock or expect it to have sympathy for my bleeding knee (regardless of if it is my pet or not) [13]. This is because the word TRUST, to a substantial amount of a population, implies something between agents with free will. It is necessary for the one whom TRUST is being placed in—let’s call them the Trustee—to have obligations towards the person who is TRUSTING them—the Trustor. One prominent theory for this belief is that these obligations are motivated by goodwill—that is, they are concerned with the Trustor’s interests, without realization of personal benefit. For the Trustor, they are willing to place

their vulnerability in the hands of the Trustee—in other words, both know the Trustee has the option to harm the Trustor and (hopefully) is choosing not to. Additionally, these roles can be reversed, as TRUST is a two-way street. If this is the case, both Trustee and Trustor must possess an emotive state and must both be full moral agents with free will, who understand their responsibilities to each other [7, 13].

Given this explanation, inanimate artefacts cannot be TRUSTED, no matter how “intelligent” they are [14]. To extend this analogy to AI, let us replace the pet rock example with an example about ChatGPT instead.

Example: I trust ChatGPT to give me the correct answers for my problem set.

If I use these answers for my problem set and discover later that my TA has given me a 50 percent on it, it would be unreasonable to feel betrayed by ChatGPT as it would also be unreasonable to feel betrayed by Google³. However, that doesn’t seem to be the case for some. There exists the phenomenon of users believing an LLM is an unbiased agent [9, 17]. However, it is neither unbiased nor agentic⁴. Suppose pilot Joe is flying a commercial airplane. There would be no negative repercussions or even any difference if we were to swap pilot Joe out for pilot Jill, because we TRUST pilots due to their training, qualifications, and shared motivation to survive the flight, rather than a personal relationship with a specific pilot. However, there would certainly be a difference if we were to swap pilot Joe out for a probabilistic automation system. Unlike a human, this system has no personal investment in landing the airplane safely, for it has no true existence or life to value, or moral consideration for the passengers’ lives [18]. Therefore, given that an LLM is not an agent and not unbiased, blindly TRUSTING that they have all the capability of a human being and none of the flaws is a mistake [6, 14, 19, 20]. Problems such as users believing they are in a relationship with their LLM, taking unsanctioned therapy advice, or thinking LLMs to be “magic” are aided by this belief that AI in its current state is a superhuman being.

To correct this assumption, a brief explanation is necessary. Current forms of AI—such as LLMs and other variants that the general public associates the term “AI” with nowadays—are ultimately very fast, mostly accurate pattern-matchers. LLMs such as Gemini, ChatGPT, and Claude are not yet capable of generating original output; rather, they are trained on vast amounts of data and are capable of rapidly recognizing the most likely subsequent pattern from an input the user gives it [21]. If you asked ChatGPT to generate a paragraph on the importance of democracy, it would select the most likely beginning word based on its database of other essays on the importance of democracy. Using that first word, it would then generate the next word, and the next, based on the most logical and common patterns. This can in no way be called “intelligence” of a higher form.

Let us take an example of someone attempting to login to their Google account. Suppose the website returns a “password incorrect” message. Colloquially, one might say “Google thinks

³As will be elucidated in more depth later on, there are cases where it is fine to ascribe the emotive states, intents, and beliefs of the creators to the product itself; in other words, a kind of shorthand. The issue is not so much that people are using this shorthand, but that they do not even recognize that it *is* a shorthand. They instead ascribe all the capacity for emotion and motivation to the product itself—in this case, ChatGPT—rather than the engineers and designers behind it. Therefore, when they say “ChatGPT believes...”, they really do think the belief stems from ChatGPT itself and not the developer who integrated this “belief” into its architecture when training it on a certain dataset.

⁴There are some that may pose the argument that LLMs as they are today *truly* are agentic. However, possibilities of their evolving behavior, which might eventually differ from hard-coded workflows, are not what we will explore now. Rather, it is necessary to proceed with the presupposition that potential for “truly” agentic AI exists in the distant future, if at all.

my password is incorrect.” Of course, we would consider the possibility that Google possesses the capacity to “think” at all to be implausible, because it has no agentic capability. What is occurring here is this: Google utilizes their comparison algorithm, which processes your password input and compares it to their stored password value. If it detects an inequality, Google will print this pre-recorded “password incorrect” message. However, we would be mistaken to think that this password comparison algorithm is something other than a program tasked with a certain objective—that is, it does not possess what we conceive to be “intelligence”.

In a more consequential example, suppose we have a machine learning algorithm that allows a doctor to better detect in a mammogram whether their patient has breast cancer. We understand that this algorithm has been trained on a set of annotated images that allow it to make a determination on a classification for breast cancer, or no breast cancer. Therefore, we view it as a “tool” or “calculator” in the hands of the doctor. On the other hand, if we begin to perceive that the algorithm is the “operator” rather than the doctor, which is the case with many conceptions of LLMs, it is more likely we will adopt a notion of its supposed “agency”, and then slip into the fallacy of TRUST.

Given that we categorize AIs as non-human systems, it would be best to evaluate whether TRUST really is the term of art to be used. Perhaps it would be better to propose an alternative: RELIANCE. The doctor could RELY on the AI algorithm to assist with a breast cancer diagnosis, but ultimately, the doctor is the accountable agent, not the algorithm. If I choose an easily guessable password for my Google account and it happens to be stolen, it would be irrational to get angry with the Google algorithm. Similarly, it would be just as irrational to get angry with ChatGPT if it gave me the wrong answers that lead to a low grade on my problem set⁵. However, this is not to say that people do not become irritated with or act irrationally towards machines. The primary issue is that they do not understand that an LLM is not sentient, and therefore incapable of responding to expressed anger. In other words, they believe it *is* rational to be angry, since the LLM is “alive” and could respond to their reaction as a human would.

III RELIANCE

Much of the literature that involves the discussion of RELIANCE defines it as what it is not, rather than what it is. Some works focus on separating TRUST from RELIANCE; when doing so, they specify exactly what constitutes TRUST, commonly stating that it is a type or form of RELIANCE [6, 13]. Given that there is extensive discourse on the definition of TRUST but few on RELIANCE, it is necessary to establish parameters in this paper for RELIANCE. In his 1994 paper, Richard Holton argues for the following definition [25]:

...to rely on someone to X is to act on the supposition that she will X: this can fall short of believing that she will X, though it is incompatible with outright belief that she will not X.

In other words, RELIANCE does not necessarily stem from “merely believing”, though it could be defined as a more practical offshoot of belief [26, 27]. Justifying why you rely on something or someone does not necessarily have to be rooted only in knowledge of prior events or rationalism (though these are, of course, helpful), but perhaps, as we will see in a later example, can be drawn from necessity, or active choice. Furthermore, RELIANCE must have some realistic grounds, even if it is something like “I didn’t have time; thus, I relied on someone

⁵This is heavily debated in the field of AI ethics. Some ethicists may disagree with this position. While it is not in the scope of this paper to discuss these views, it may be interesting for future research [22, 23, 24]

else to complete the task.” Belief, on the other hand, could stem from pure epistemics, and does not have to be actively chosen [25].

If we take this definition to be a well-specified and complete one, we can further distinguish RELIANCE to be an attitude not influenced by irrational factors (or extra factors that distinguish it from TRUST), or anthropomorphic influences [14, 27]. Furthermore, we have already established that TRUST requires both parties to be full moral agents (which necessitates the Trustee and Trustor to also possess wills and emotive states); RELIANCE, on the other hand, does not by definition. RELIANCE also lacks an anthropomorphic component. It is more accurate to say, “I rely on my heater to warm my house,” rather than “I TRUST my heater to warm my house.” Therefore, RELIANCE would be both a more accurate and less dangerous term of art to use for AI. Below are some examples of accurate usage of reliance:

I rely on my car to get me to the diner.

I rely on my Wi-Fi connection to not cut out.

I rely on ChatGPT to provide me with solutions to the problem set.

I rely on my hammer as a means to drive in nails while building a table.

One of the main distinguishing factors is that unlike TRUST, RELIANCE has less normative value attached to it. Let’s say that I am planning a party. I decide to rely on my friend Amy to buy balloons for the party, because I have too many other things I have to do. There is no evidence that I should believe Amy will not buy the balloons for me. However, suppose that in the past Amy has occasionally not done the tasks I have set for her; in this case, there is also not absolute belief that she certainly will go and buy the balloons. However, even if Amy does not buy the balloons, no injury or betrayal will emerge from this. I might be disappointed, but there would not be any harm to our fundamental relationship. On the other hand, when you TRUST someone, there will always exist the possibility that your Trustee will injure you [13]. This is not necessarily a condition of RELYING on another person or object, as it is unlikely a reasonable person would assume that someone who, or even something that, they rely on owes them anything in return. For example, assume we have a case in which I tell Amy that one of my relatives is going through a divorce, but that she should keep it a secret. If she then spreads this information to our other friends and coworkers, I would certainly feel hurt by her actions, as I TRUSTED her to not do that. In this case, TRUST is the correct word to use in this situation, because I not only RELIED on her to not say anything, but also felt safe enough to be vulnerable. In this case, she broke my confidence and caused me emotional pain.

A word of caution here. Some accounts use RELIANCE and “reliability” interchangeably when referring to AI [28]. This is another error. “Reliability” carries connotations of measurable and numerical properties—stemming from engineering and technological disciplines—which we hope to avoid. RELIANCE lacks these unwanted associations. Hence, RELY and RELIANCE should not be grouped with “reliability”. “Reliability”, or even “reliable [AI]” should also not be used as a term of art here⁶.

⁶There are cases in which “reliable” can be applied to certain services, even with AI. However, this would be applicable more for engineering use cases, rather than common ones. For example, it would be alright to say that an LLM could “reliably” hit certain targets they have been engineered for, such as giving an answer to a prompt 100 percent of the time. However, this is different from giving the correct answer. If asked a question outside their set of training data, an LLM would still provide an answer and deliver it with confidence (in order to hit the aforementioned engineering target), whereas a human might tell you they would not be able to give an answer. Therefore, we would be able to RELY on a human to give the correct answer—even if that means saying “I don’t know,”—but not on an LLM, as they would be hitting their engineering targets “reliably” by providing an answer, which itself is unreliable.

Similarly, some may wonder if it is acceptable to use “reliance-worthy” as a replacement for “trustworthy” when referring to AI. Again, this is not an acceptable alternative. Although we do say things such as “This is a reliable hammer”, or “This is a reliable car”, these are first, shortcuts, and second, a significant amount of the population would never assume that a qualifier of “reliable” when used on a hammer or car ascribes an agentic capacity to them. However, as we have already explained, there exists a widespread belief that current AIs have agency, and thus can be trustworthy [29]. Since we want to avoid propagating this belief further, we will not use a term such as “reliance-worthy AI” but instead adhere to phrases that maintain full agency and decision-making with ourselves, the humans, such as:

I rely on the AI.

The AI can be relied upon.

I believe reliance on this AI is a wise choice.

IV Does Safety Assure RELIANCE?

Assuring safety is presuppositional to relying on a system. There are many methods of safety assurance, and some, such as ones developed for the avionics industry, may be well-suited to application in the AI realm [30, 31]. Nevertheless, we still must deal with the gap between safety and RELIANCE. In other words, can we be sure that a safe system is one that we can RELY on with minimal remaining doubt?

To answer the question posed in this section, comparisons that can be analogues between the relationship of safety and reliability, and safety and RELIANCE, will be used⁷. For example, high levels of reliability are sometimes correlated with high levels of safety, but this is not always the case. Sometimes, increasing reliability of an artefact or system leads to decreasing levels of safety; other times, the opposite is true. Let us say we are dealing with a nuclear reactor system. If we never turned the system on, it would be very safe, but we could not call it reliable. On the other hand, let us say the system functioned reliably to generate electricity, but emitted radiation that killed the nearest worker every time it was turned on. This is obviously unsafe, although reliable in its function. Following a similar line of reasoning, we can posit that high safety does not necessarily imply a high conviction in our ability to RELY on, or to have high RELIANCE in, a system.

Although we hope to avoid connotations of measurability or numerical properties by using the term RELIANCE, it is beneficial to introduce the concept of measurable, or modeled, reliability, as another helpful analogue for a problem we may run into with RELIANCE. That is, we should not always RELY on a system or artefact merely because it has high modeled reliability. Modeled reliability only considers properties that can be measured, or properties that people see fit to introduce into the model, thereby overlooking a whole host of factors that either cannot be modeled or were never considered. This leads to two major consequences. First, due to this elimination of “unnecessary” or “extraneous” factors, modeled reliability cannot be a complete metric. Therefore, users should not mistakenly over-RELY on systems or artefacts when they see it possesses high modeled reliability. Second, users may also not understand that they are relying on these systems or artefacts in an incorrect manner or may

⁷Although we are using the safety-reliability relationship as an analogue for comparable situations for safety-RELIANCE, it is still important to remember that reliability should not be and is not being used interchangeably with RELIANCE.

assume—mistakenly—that a future situation will always produce the same result as a past situation. A person using a driver-assist system in the car may have had a situation (or several) in the past when their car helped them brake in time. However, it would be mistaken to be perpetually assured that the car will brake every single time. Perhaps the person also relies on the brake assist in situations that the system was not built for, due to their assumption that it would apply to every possible scenario. Additionally, suppose this driver-assist system is reliable 98 percent of the time. There is no random collection of events where the 2 percent of failure shows up; rather, it happens in specific situations. Suppose this failure only shows up when a child runs in front of a car. A 2 percent chance of failure—which might be called a small number—does not assure us of its safety, or reliability. We must ask ourselves where the outliers in the sample space are, and furthermore, what these outliers lead to. If these outliers only lead to a faulty hazard light, it would be more reassuring to rely on this system. If these outliers, as in the previous example, lead to injuring a child, RELIANCE on this system is not possible. Thus, modeled reliability cannot inherently assure that one should apply their RELIANCE to a situation.

If modeled reliability, safety assurance, and measurable properties are not all that completely assure that we can RELY on something, what other factors could be involved? Let us introduce another thought experiment to determine a possible answer to this question. Say we have an alarm system in our home that detects if somebody is attempting to come in through the front door. Say we also know this alarm has a confirmed modeled reliability (not RELIANCE) of 99.5 percent. However, if the alarm goes off and we decide to check if a burglar is at the front door, we are not conducting a reliability test. Nor would most of the population be actively recalling that “This alarm has a reliability of 99.5 percent; therefore, I will decide to rely on it.” When the moment of RELIANCE comes, it is not based off only epistemics, and neither can it be called a purely active decision. Reasons of practicality and reinforced behavior in the presence of past responses to alarms also affect the moment of RELIANCE.

Nevertheless, there is clearly a gap between concrete testing, such as in safety assurance, and measurable properties such as reliability, and the moment at which we decide to act upon this. We have a missing link between the so-called “real world” and the world of abstract things, such as knowledge and perception. This is called the epistemic gap in philosophy [32]; for our purposes, it may be helpful to refer to it as the uncertainty gap.

Let us construct a thought experiment to work through this question. Assume we create a deterministic model to measure the amount of fuel we have, using sensors in the aircraft’s tanks for detection. The assumptions it uses are ones that the industry has used in the past. Everything about this system builds on an iteration that previously has been executed and tested; only the deterministic model itself is completely new. Then also assume another person is working on the same task. Instead of a deterministic model, they are creating a probabilistic model to measure the amount of fuel in the tanks. This model does not use the same assumptions as our deterministic model, and neither does it build on things that the industry has done before. Unlike with the deterministic case, where only the model itself is a new and untested ‘version’, the probabilistic model has no precedent, in both assumptions and structure. In other words, we are concerned with whether the deterministic model will measure fuel correctly under the systems and assumptions we use. With the probabilistic model, we are concerned with whether this novel method *itself* is the correct process. A crude simplification for our moment of RELIANCE on this probabilistic model could be akin to this: the point at which we choose to RELY on the deterministic model requires one step over the uncertainty gap, whereas the probabilistic model requires two—or some number greater than one—steps over the uncertainty gap.

If we extend this thought experiment to AI, it is reasonable to conclude that there would be a massive uncertainty gap to overcome. Not only is the nature of the models we are using probabilistic, which inherently carries uncertainty, but there is also not nearly enough evidence or historical backing in the field upon which we can build. This contrasts with the field of avionics, where there have been long decades of prior testing and constant reassessment [33]. With avionics, introducing new systems or artefacts can be compared to a short step over an uncertainty gap of known distance. With AI, choosing to rely on it is a leap; simultaneously, the dimension of its uncertainty is so far unknown.

V Warnings

Someone who does not believe exploring the concepts in this paper is worthwhile, may offer rebuttals to the value such as the following:

1. It is pointless to argue about semantics.
2. The term is already well-established in the research community and wider circles. Attempts to dismantle this paradigm, and others like it, is not worth the effort given the minimal gain.
3. Trust is a good thing.

In this section, we will attempt to present counter-rebuttals against these three points. These are summarized as follows (ordered sequentially based on the arguments above):

1. The only thing worth arguing about is semantics.
2. Dismantling this paradigm does not lead to minimal gain. In fact, the work to dismantle it is very important considering how people beyond the research community are handling AI right now.
3. The population of people who believe that trust is a good thing fall into three categories, which we will discuss separately:
 - (a) They do not yet understand the harm that emerges from using terms such as trust.
 - (b) They are acting in bad faith.
 - (c) They truly believe that encouraging trust in AI is a good thing.

Rebuttal 1 is easy to refute. Semantics concerns how we give meaning to words. Although there may be bad faith arguments made using semantics to deliberately obfuscate the substance of the original discussion, this is not what we are attempting to do with our distinctions. Some of the most important and controversial arguments in philosophy deal with defining terms; for example, the differences between ‘killing’ and ‘letting die’ [34]. By being precise about the correlations and implications TRUST carries, versus those of RELIANCE, we can begin to establish definitions that are truly appropriate for usage in these contexts, especially since so many of these implications can be detrimental to our goal.

Rebuttal 2 comes from a sector who: 1) doubt it is possible to change usage of “TRUST in AI” to “RELIANCE in AI”, and 2) believe that regardless of these attempts, it would not be worth the time or effort to change this paradigm [15]. However, it is both possible and

worth the effort. One of the dangers currently circulating in the AI industry are attempts by companies and investors to increase TRUST by manufactured and marketed methods. This can be called the “commodification of TRUST”, a phenomenon in which practices such as anthropomorphization are promoted to increase a user’s TRUST—in this case, in AI—to benefit from it. If people TRUST AI more, it follows that they are more willing to spend money, use those models, and support the company [1, 20, 35]. Research has found that these efforts are successful, and have extended not just to the general public but the research community [15], demonstrating how the constant and deliberate usage of anthropomorphic terminology only furthers over-RELIANCE on AI. Given that this is a widespread and increasingly prevalent threat, it would be both defeatist and inaccurate to say that it is not worth the effort of changing the terms used for AI.

Rebuttal 3 is a general statement that, upon closer inspection, stems from three broad groups. Group 3(a) says it because they are not aware at all, or not aware to the fullest extent, of the harm in using words like TRUST to refer to AI. However, it is our hope that this paper will shed light on its detrimental effects and allow people to think more critically about their usage of the word TRUST, and, if they would prefer to continue using it, to have been made aware of its possible signals and implications.

Group 3(b) says that TRUST is a good thing because they have other motives for encouraging TRUST in AI. These motives are often malicious. They could be doing it to commodify TRUST, which, as we have already elucidated, is not a good thing. They could also be doing it because they believe it offers some kind of personal benefit through ethicswashing their motives [36]. Whatever the case, we cannot accept this as a valid rebuttal given the way they seek to weaponize TRUST.

Group 3(c) says that TRUST is a good thing because they truly believe that we will both create AIs that are truly agentic and possessed of free will, and that we should put our TRUST in them when these AIs emerge [37]. We can also call this the ‘good faith’ argument. While there exists extensive discourse (and belief) concerning the emergence of agentic AI, it is not in the scope of our discussion to delve into these arguments, whether possible or not. Neither is it in our purview to discuss if we should TRUST agentic AI. To the third group, we will merely say that truly agentic AI does not exist yet, and that it is a mistake to believe any of current LLMs or other forms of AI, generative or otherwise, that are used and developed today, are near the capability of having free will or agency [38]. Therefore, their second proposition that TRUST in AI is desirable cannot yet be valid.

Finally, beyond concerns around shifting the TRUST paradigm to a RELIANCE paradigm, there is another point of warning necessary to make. Though we have discussed in length the dangers of anthropomorphizing AI and using anthropomorphic terms to refer to them, this poses the threat of shifting too far in the opposite direction: thinking of AIs as nearly completely akin to calculators⁸. A calculator follows a structured strategy and algorithm. We have minimal remaining doubt in its output because we can follow a series of well-understood principles that underpin its calculations. However, it is well-documented that humans are more likely to over-RELY on inanimate systems, such as AI, because of the mistaken belief that these systems are free of human bias and/or are infallible⁹ [39]. This gives AIs such as LLMs undeserved credence; we implicitly assume its output must be correct. However, LLMs are not thoroughly understood, even to this day; the black-box problem is well-documented and still being worked on despite recent breakthroughs [40, 41]. Additionally, regardless of the

⁸Calculators here is being used as a general term; we could also say it is being thought of as an explainable, inanimate tool.

⁹Whereas we understand and are cognizant of human fallibility.

in-explainability of its computing process, it is still dangerous to assume the system will never fail simply because it appears as though a human is not making decisions¹⁰ [43]. Therefore, just as we must be cautious to avoid anthropomorphizing AI, so too must we be aware of the dangers in treating them as simple calculators we can consistently RELY on.

VI Conclusion

In a world where AI is increasingly used in all manners from self-driving cars, to large language models, to making medical breakthroughs, it is necessary to delve deeper into the way we think about these models. Setting forth an alternative to the paradigms used today may not only encourage deeper consideration about the true suitability of TRUST in relation to AI, but also introduce caution where these paradigms are being used extensively. This paper introduced the concept of RELIANCE as a more informed and cautionary manner of dealing with AI. Several informal examples were used to elucidate the relationship between TRUST and RELIANCE, the usage of RELIANCE itself, and the relationship between safety and RELIANCE. In future research, it would be interesting to conduct a more formalized thought experiment utilizing safety assurance methodologies to elucidate how uncertainty gaps might be reduced to better RELY on systems. Stemming from this, it would also be beneficial to elucidate if creating an AI we can RELY upon is currently possible, or a wise choice.

Acknowledgments

This work was completed during the Summer 2026 session of the NASA Langley Research Center Office of Science, Technology, Engineering, and Mathematics (OSTEM) Internship program under the expert mentorship of C. Michael Holloway.

I would like to sincerely thank Taylor A. Thoman, Mallory Graydon, PhD., Sarah Lehman, PhD., and Aaron Dutle, PhD., for their insights, incisive revisions, and helpful feedback.

To Reach the Author

Please send e-mail to `victoria.tan@yale.edu` and `c.michael.holloway@nasa.gov`.

References

- [1] Inie, N.; Zukerman, P.; and Bender, E. M.: De-anthropomorphizing AI: From Wishful Mnemonics to Accurate Nomenclature. *First Monday*, 2026. URL <https://firstmonday.org/ojs/index.php/fm/article/view/14366/12152>.
- [2] Noorbehhahani, F.; and Oyibo, K.: AI Over-dependence and Human Cognitive Decline: Hazards, Evidence, and Mitigation Strategies. *Computers in Human Behavior Reports*, 2026, pp. 101–102.
- [3] Killoran, J.; and Park, A.: Learn from the Blame Game when AI Causes Harm. *Proceedings of the National Academy of Sciences*, vol. 123, no. 17, 2026.

¹⁰See the incisive quote by Koopman: “Even if you do a billion miles of simulation, how do you know the simulation software and models are essentially perfect?” [42]

- [4] Lee, J. D.; and See, K. A.: Trust in Automation: Designing for Appropriate Reliance. *Human factors*, vol. 46, no. 1, 2004, pp. 50–80.
- [5] Baier, A.: Trust and Antitrust. *ethics*, vol. 96, no. 2, 1986, pp. 231–260.
- [6] Ryan, M.: In AI we Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, vol. 26, no. 5, 2020, pp. 2749–2767.
- [7] Sutrop, M.: Should We Trust Artificial Intelligence? *Trames*, vol. 23, no. 4, 2019, pp. 499–522.
- [8] Everett, J. A.; Claessens, S.; Knöchel, T.-D.; and Reinecke, M. G.: Principles for Understanding Trust in Artificial Intelligence. *Nature Reviews Psychology*, 2026, pp. 1–14.
- [9] Kazim, E.; and Koshiyama, A. S.: A High-Level Overview of AI Ethics. *Patterns*, vol. 2, no. 9, 2021.
- [10] Henrique, B. M.; and Santos Jr, E.: Trust in Artificial Intelligence: Literature Review and Main Path Analysis. *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 1, 2024, p. 100043.
- [11] Afroogh, S.; Akbari, A.; Malone, E.; Kargar, M.; and Alambeigi, H.: Trust in AI: Progress, Challenges, and Future Directions. *Humanities and Social Sciences Communications*, vol. 11, no. 1, 2024, p. 1568.
- [12] Duenser, A.; and Douglas, D. M.: Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust. *IEEE Intelligent Systems*, vol. 38, no. 6, 2023, pp. 19–26.
- [13] Hawley, K.: Trust, Distrust and Commitment. *Noûs*, vol. 48, no. 1, 2014, pp. 1–20.
- [14] Landes, E.; Claessens, S.; and Everett, J. A.: Trust in AI is not a Conceptual Confusion. *Preprint*, March 2026. URL <https://philpapers.org/archive/LANTIA-8.pdf>.
- [15] Freiman, O.: Making Sense of the Conceptual Nonsense ‘Trustworthy AI’. *AI and Ethics*, vol. 3, no. 4, 2023, pp. 1351–1360.
- [16] McLeod, C.: Trust. 2023. URL <https://plato.stanford.edu/archives/fall2023/entries/trust/>.
- [17] Nissenbaum, H.: Accountability in a Computerized Society. *Science and Engineering Ethics*, vol. 2, no. 1, 1996, pp. 25–42.
- [18] Milford, S. R.: The Last Human Touch: Why AI Doesn’t Belong in Life’s Defining Moments. *BMC Medical Ethics*, vol. 27, no. 1, 2026, p. 98.
- [19] Viehoff, J.: Making Trust Safe for AI? Non-agential Trust as a Conceptual Engineering Problem. *Philosophy & Technology*, vol. 36, no. 4, 2023, p. 64.
- [20] McMillan, D.; and Brown, B.: Against Ethical AI. *Proceedings of the Halfway to the Future Symposium 2019*, 2019, pp. 1–3.
- [21] Raza, M.; Jahangir, Z.; Riaz, M. B.; Saeed, M. J.; and Sattar, M. A.: Industrial Applications of Large Language Models. *Scientific Reports*, vol. 15, no. 1, 2025, p. 13755.

- [22] Rees, C.; and Müller, B.: All that Glitters is not Gold: Trustworthy and Ethical AI Principles. *AI and Ethics*, vol. 3, no. 4, 2023, pp. 1241–1254.
- [23] Bender, E. M.: Artificial Intelligence. *Oxford Research Encyclopedia of Science, Technology, and Society*, R. Fouché, ed., Oxford University Press, 2026. URL <https://doi.org/10.1093/9780197852712.003.0067>.
- [24] Weil, G.: Abnormally Dangerous Algorithms: The Case for Strict Liability at the AI Frontier. 2026. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6588958.
- [25] Holton, R.: Deciding to Trust, Coming to Believe. *Australasian Journal of Philosophy*, vol. 72, no. 1, 1994, pp. 63–76.
- [26] Alonso, F. M.: What is Reliance? *Canadian Journal of Philosophy*, vol. 44, no. 2, 2014, pp. 163–183.
- [27] Alonso, F. M.: Reasons for Reliance. *Ethics*, vol. 126, no. 2, 2016, pp. 311–338.
- [28] Durán, J. M.; and Pozzi, G.: Trust and Trustworthiness in AI: Durán and Pozzi. *Philosophy & Technology*, vol. 38, no. 1, 2025, p. 16.
- [29] Zanotti, G.: AI Systems Should be Trustworthy, not Trusted. *AI & SOCIETY*, vol. 41, no. 4, 2026, pp. 3401–3412.
- [30] Holloway, C. M.: Understanding the Overarching Properties. Technical Memorandum NASA/TM-2019-219650, National Aeronautics and Space Administration, Hampton, VA, USA, July 2019. URL <https://hdl.handle.net/2060/20190029284>.
- [31] Thoman, T.: Extending the Overarching Properties with Ethical Principles. White paper, OSTEM Internship Program, NASA Langley Research Center, Hampton, VA, USA, June 2026. URL <https://ntrs.nasa.gov/citations/20260004495>.
- [32] Metz, J.: Omissions, Moral Luck, and Minding the (Epistemic) Gap. *Canadian Journal of Philosophy*, vol. 51, no. 4, 2021, pp. 301–314.
- [33] Boulter, D. H.: Roadmap for Artificial Intelligence Safety Assurance. *Aviation Safety Report. Federal Aviation Administration*, 2024.
- [34] Räsänen, J.; and Häyry, M.: Fifty Years of Killing and Letting Die: On the Limits of Philosophical Bioethics. *Bioethics*, vol. 40, no. 3, 2026, pp. 319–323.
- [35] Krüger, S.; and Wilson, C.: The Problem with Trust: on the Discursive Commodification of Trust in AI. *Ai & Society*, vol. 38, no. 4, 2023, pp. 1753–1761.
- [36] Schultz, M. D.; Conti, L. G.; and Seele, P.: Digital Ethicswashing: A Systematic Review and a Process-Perception-Outcome Framework. *AI and Ethics*, vol. 5, no. 2, 2025, pp. 805–818.
- [37] Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B.: Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, vol. 55, no. 9, 2023, pp. 1–46.
- [38] Hauhio, I.; Buder-Gröndahl, T.; and Wallenberg, A.-M.: Do AI Agents Have Free Will, Really? *AI and Ethics*, vol. 6, no. 3, 2026, p. 331. URL <https://doi.org/10.1007/s43681-026-01125-z>.

- [39] Klingbeil, A.; Grützner, C.; and Schreck, P.: Trust and Reliance on AI — An Experimental Study on the Extent and Costs of overreliance on AI. *Computers in Human Behavior*, vol. 160, 2024, p. 108352.
- [40] Asatiani, A.; Malo, P.; Nagbøl, P. R.; Penttinen, E.; Rinta-Kahila, T.; and Salovaara, A.: Challenges of Explaining the Behavior of Black-box AI Systems. *MIS Quarterly Executive*, vol. 19, no. 4, 2020, pp. 259–278.
- [41] Komera, O.; and Manche, R.: Black-box Behavior in Large Language Models: Challenges and Implications. *Proceedings of the International Conference on Renewable Energy, Green Computing and Sustainable Development*, 2023.
- [42] Koopman, P.: UL 4600: What to Include in an Autonomous Vehicle Safety Case. *Computer*, vol. 56, no. 5, 2023, pp. 101–104.
- [43] Salatino, A.; Prével, A.; Caspar, E.; and Lo Bue, S.: Influence of AI Behavior on Human Moral Decisions, Agency, and Responsibility. *Scientific Reports*, vol. 15, no. 1, 2025, p. 12329.