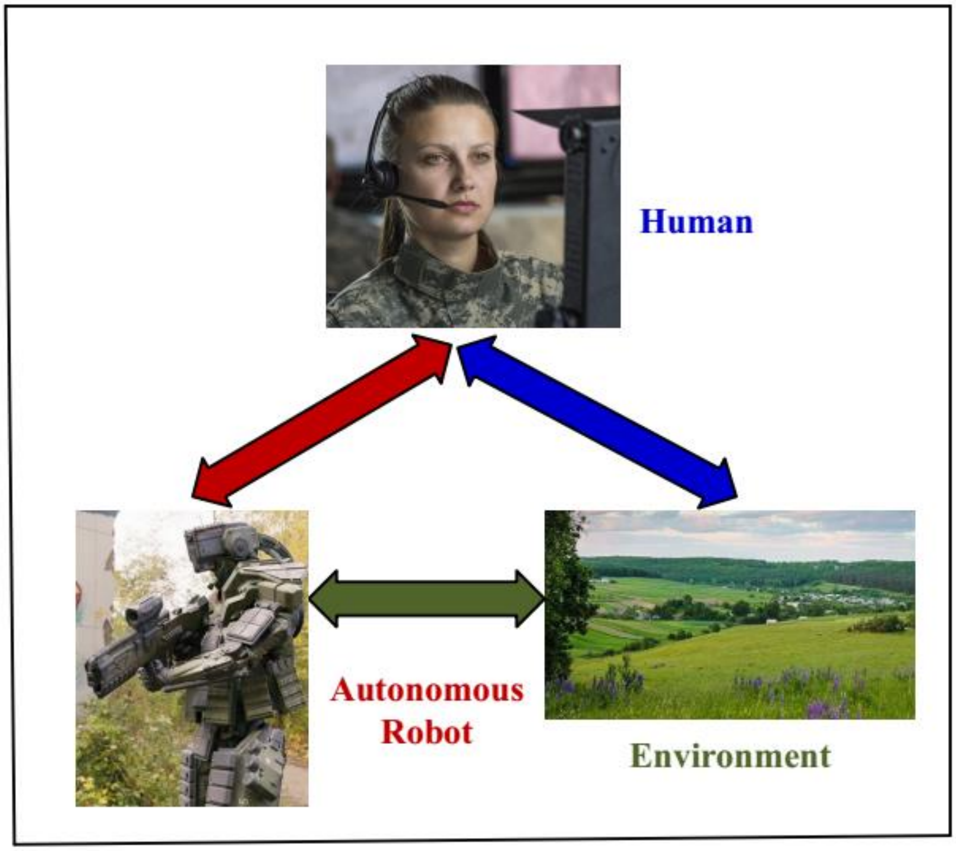
White Paper
Suggested Metrics for Trusted Autonomy

Dr. Robert Finkelstein, President
Robotic Technology Inc.
BobF@RoboticTechnologyInc.com
Office: 301-983-4194



3 January 24

Trusted Autonomy: *the basis and process for establishing trust in autonomous systems.*

1.0 Purpose:

The purpose of this White Paper is to suggest and define metrics for trusted autonomy. Autonomy is increasingly sought globally for artificial intelligence (AI), robots, robotic vehicles (i.e., robotic air, ground, and water vehicles), and other systems for military and civil applications, and the technology for autonomy is rapidly becoming feasible for unconstrained environments and unsupervised operations. With this growing demand for autonomy soon to be satisfied by technological advances, there is an increasing expectation for trusted autonomy to allay fears of AI overlords and a robot apocalypse.

2.0 Trust

Philosophers have contemplated the meaning of trust and trustworthiness for centuries, focused on humans, and its meaning remains nebulous. Trust is important, but also dangerous. Trust requires us to depend on others, whether humans or autonomous machines. If the others were guaranteed to provide what we want from them, we would not need to trust them. But our risk is that they will fail to provide what we want, so that our trust was misplaced.

In the context of machines, and especially autonomous machines, there is no universally accepted definition of “trust”, so that those involved with such systems tend to view trust differently. One view is that trust must be between conscious entities, such as between humans or a human and an autonomous (cognizant) machine (not merely an automaton). Mutuality requires that while a human must be able to trust an autonomous robot, the robot must also be able to trust the human. Also, trust inherently involves risk and peril. If it were known with complete certainty that an entity would accomplish what was requested, then trust would not be necessary. Trust is required only if there is some uncertainty in the accomplishment of the requested task. Despite its ambiguity, we will define trust between people and autonomous systems sufficiently for pragmatic applications.

2.1 Key Variables of Trust

The determination of trust is objective and subjective, with psychological and sociological characteristics. The levels or amount of trust forms a continuum, where trust can range from high to low. At some point, trustworthiness becomes untrustworthiness. There are several variables comprising trust, which could serve as metrics for trust. These variables or metrics are not of equal value, with some more important than others. The key variables of trust that determine the level of trust are listed below, not necessarily in order of importance, followed by a brief discussion of each, describing how they impact trusted autonomy.

- Risk and risk mitigation
- Uncertainty
- Reliability
- Accuracy and precision
- Learning and adaptation

- Redundancy
- Predictability
- Transparency
- Consistency
- Robustness and Resilience
- Security
- Situational and Self-Awareness
- Value Judgment
- Context Cognition
- Self-Reflexive Meta Control System

Risk and Risk Mitigation: If a task or function had no risk associated with its performance, then trust in the autonomous system performing would not be necessary. Military missions, however, usually have significant risk associated with them. Risk may be defined in a number of ways. But a pragmatic definition is that risk is a function of the product of the probability that an (an adverse) event will occur and the expected consequence (or impact) of the event, should it occur:

Risk = (probability of the occurrence of an event) x (consequences if the event occurs)

Risk is contingent and generally subjective, depending on whose ox is getting gored (or whose mission will catastrophically fail). It depends often on subjective determinations of both the probability of an event and the consequent impact of the event. The acceptability or unacceptability of a level of risk (and need for risk mitigation) depends on whether the decision maker (which may be an autonomous system) is risk averse or risk taking. Experiments with autonomous cars, for example, varied the levels of risk for cars at intersections with four-way stop signs. Cars that are programmed to be highly risk averse can spend excessive time waiting to cross the intersection, while cars that are too risk taking are more likely to have accidents. Trusted autonomy would adjust the acceptable risk of an autonomous system to minimize accidents (or other failure) while allowing the system to accomplish its mission.

Trusted autonomy for military missions will require the ability of the autonomous system to judge the risk involved with executing various tactics when confronted with a specific adversary within a specific environment. The system, for example, would need the expertise and situational awareness of a leader who was a combat veteran. Autonomy is more trustworthy if risk can be *mitigated*, such as by the system's ability to analyze its situational risk, its ignorance of the situation, its situational uncertainty, and then re-plan its mission to accomplish it.

Uncertainty: Uncertainty is caused by incomplete knowledge due to inherent deficiencies in knowledge that has been acquired. Uncertainty can be classified into three types based on its sources: ambiguity, approximations, and likelihood.

Ambiguity arises from possible multiple outcomes of an event, such as whether an army's planned blitzkrieg-type attack will quickly succeed or be stifled by an unexpectedly resistant defense.

Approximations arise because the human mind has the ability to perform estimates through (analytic) reduction and (synthesis) generalizations, such as employing deduction and induction, respectively, in developing knowledge. The process of approximation can involve the use of vague semantics in language, approximate reasoning, and emphasizing relevance to simplify the complexity of a situation, as in a government's propaganda to its citizens. Approximations may employ vagueness, coarseness, and simplification. *Vagueness* results from the non-crisp nature of the belonging and non-belonging of elements of a set (or a notion of interest), such as the motivation of a dictator in attacking a neighboring nation. *Coarseness* results from approximating a crisp set by subsets of an underlying partition of the set's universe that would bound the crisp set of interest, such as attributing superior decision-making capabilities to an autonomous system based on its performance on a limited array of tests and experiments. *Simplifications* are assumptions made to make problems and solutions tractable, such as generalizing about the tactical capabilities of an autonomous combat vehicle.

The *likelihood* component of uncertainty can be defined in a context analogous to that of chance, odds, and gambling, and it has the primary components of randomness and sampling. Ignorance from *randomness* stems from the unpredictability of outcomes, while ignorance from *sampling* arises from the use of limited samples of a population to characterize the entire population. As with humans dealing with life in general, autonomous systems must be able to determine the likelihood that its decisions and actions will be successful for an acceptable record of efficacy.

Trusted autonomy for military missions will require the autonomous system to be able to deal with, and sufficiently overcome, the constant web of uncertainty that pervades conflict.

Reliability: Reliability is sometimes considered a synonym for "trustworthy", although we consider "trustworthy" to have many more characteristics associated with it. The attribute of consistency is associated with reliability, especially if an entity is performing consistently well. But consistency need not imply successful performance; merely repeatable performance, providing the same results under the same conditions. An autonomous system would be trustworthy if it were to function consistently as expected, which means its performance may not be acceptable even if it were consistent. If, for example, an autonomous vehicle's speed were consistently 5 miles per hour slower than its actual speed, it would arrive at a certain objective in 2 hours instead of 1.5 hours. It would be reliably late – its late arrival precisely known in advance. Its speed could be adjusted to account for the error, or such a consistent error could be accounted for in planning the mission so that the mission would result in success despite the consistent error.

Nevertheless, we will assume a canonical requirement for acceptable reliability that the performance of a trusted autonomous system (as with any system) has been successfully verified against its designer's specifications and validated against its user's expectations, and that the cause of any deviations can be or have been repaired.

Accuracy and Precision: Accuracy and precision are measures of error, with the former determining the difference between measurements and their true value, and the latter determining the dispersion among repeated measurements, e.g., a measure of variability or random errors. To be trusted, autonomous systems must be acceptably accurate and precise for their purpose or function. An autonomous vehicle, for example, must have the ability to brake with acceptable accuracy and precision. If it is expected to stop within 40 feet at 20 mph, then it must do so repeatedly. Likewise, all of the functions of a military robot must be performed according to its specifications (within an acceptable margin of error) every time; otherwise it will not be trusted. No one wants be a passenger in an autonomous vehicle that, when confronted with an obstacle, sometimes stops within 80 feet at 20 mph, instead of within 40 feet.

Learning and Adaptation: Learning is the acquisition of knowledge, skill, ability, or understanding by study, instruction, or experience, as evidenced by achieving growing success (improved behavior), with respect to suitable metrics, in a fixed environment. Learning takes place when the autonomous system's behavior increases the efficiency with which data, information, and knowledge is processed so that desirable states are reached, errors avoided, or a portion of the system's environment is controlled. Adaptation is a change in an autonomous system's behavior (or structure) in response to a changed environment. The system is able to maintain critical or essential variables within physical (or physiological) limits (e.g., homeostasis), and the system's changed behavior (or physical structure) increases the probability that the system can achieve its function or purpose by adjusting to the new or changed environment. To summarize: we make a distinction between learning and adaptation, where Learning takes place in a *fixed* environment, and adaptation takes place in a *changed* environment. The trust in an autonomous system is enhanced if the system is able to learn from its experiences and adapt to changes in its situation or environment. For example, an autonomous combat robot that is able to learn to distinguish among newly encountered civilian vehicles, friendly military vehicles, and enemy military vehicles, and can also adapt to changing enemy tactics, is more trustworthy than one not able to do so.

Redundancy: In a system a redundant component or subsystem is a replica of an existing component or subsystem, implying that it is unnecessary and inefficient, taking up space, increasing weight, and adding to cost. But sometimes efficiency is better sacrificed for effectiveness, to ensure the system is able to carry out is mission or perform its function despite degradation to its physical structure or operational performance, to enable graceful degradation. If the main brakes fail, it is useful to have another means of stopping, e.g., *emergency* brakes. An autonomous system cannot have a redundant version of every component – it is not practical. But if certain critical or especially vulnerable components had backup versions, the autonomous robot would be more survivable on the battlefield, for example, able to carry out its mission with a greater probability of success. Alternatively, swarms of (inexpensive) robots with redundant capabilities could, en masse, better survive combat attrition to accomplish the mission. Thus trusted autonomy increases with (appropriate) robot redundancy.

Predictability: A predictable system behaves as expected. The system is not only consistent in its behavior, but we know in advance what its behavior will be, the opposite of que será, será (whatever will be, will be). This system attribute is usually a good thing, to know that pressing the brake a certain amount causes the car to decelerate at a certain amount and a stop at a predictable distance. The predictability of the behavior of the system's components provides confidence – trust – in the system. However, the predictability of the operational or tactical behavior of an autonomous combat robot or vehicle can be fatal if the adversary is the one doing the predicting. Therefore, predictability of an autonomous system should be concerned with its ability to make appropriate decisions and behave correctly to carry out its missions successfully. The actual decisions and subsequent details of the tactical behavior of the robot should not always be predictable, at least not by an adversary. For example, it may be predictable that a platoon of autonomous battlefield robots will decide to ambush an adversary. But the chosen location of the ambush, and the type of cover and concealment preferred by the robots in the context of the terrain, may be sufficiently varied and creative as to be unpredictable. Trust in the robot platoon's autonomy arises from the expectation that each robot in the platoon will perform its function as expected, and that the robot platoon's tactics will be successful.

Consistency: For the systems of interest, consistency means that they always behave in the same way. The autonomous robot that is considered consistent, however, is not expected to literally behave identically in every instance, like a properly functioning elevator. Rather, its consistency is means that its components and subsystems each perform within their expected specifications, and that its decisions are correct within the context of the various situations confronting the robot, i.e., its decisions and consequent (tactical) behavior are the right ones even though they are not necessarily repeatable. This sort of consistency enhances trust in the autonomous robot, whether form military, civil transportation, or household applications.

Transparency: Transparency has become a major issue in artificial intelligence because some of its tools and techniques, such as deep learning (neural networks), are notoriously not transparent. Transparency here means the ability of a human to see why an autonomous system made a particular decision or behaved in certain way. An autonomous system must be able to explain and justify its decisions and behavior. Rule-based (or expert) AI systems, for example can explain their behavior, e.g., "I did this because of these conditions and these rules relevant to the conditions". The human mind is rarely transparent, but we have more or less accommodated to humans and their behavior regardless of transparency. An opaque autonomous machine system is not conducive to being trusted, leading to pessimistic visions of Skynet and the Terminator. Transparency engenders trust in an autonomous robot, and there are approaches to achieving transparency with reflexive processes, discussed below.

Robustness and Resilience: The *robustness* of an autonomous system allows it to withstand or overcome adverse conditions, allowing it to perform successfully under various hostile conditions. If the system fails, it degrades gracefully, gradually, not catastrophically, so there is ample time to for a mission to succeed (e.g., the brakes will still work for a while despite a warning that they will soon fail). The *resilience* of an autonomous system implies that it can recover quickly from damage or failure, either by repairing itself, being repairable by humans, or adjusting to the impairment (e.g., by a work-around). An autonomous system that degrades gracefully and that can repair itself (or continue to function despite damage), is more trustworthy than a system that is not so deft at recovery.

Security: Autonomous systems, whether civil cars or military robots, must be secure from remote hacking or electronic damage (e.g., a microwave weapon). The adversary cannot be allowed to remotely take control of the autonomous system or electronically interfere with it or damage it. An autonomous system that is vulnerable to hacking or electronic interference cannot be trusted.

Situational Awareness: Having situational awareness is the ability of an autonomous system to perceive its environment and the entities in it, to understand the significance, to itself and its mission, of what it is perceiving, and to anticipate the near-term future of the environment and the entities in it relative to itself and its mission. The military's OODA Loop (Observe, Orient, Decide, and Act) is three-quarters relevant to situational awareness; only *Act* is not relevant, because any action is subsequent to becoming aware of the situation. The autonomous systems gathers information about its environment, analyzes and understands the information, predicts how the environment might change during its mission, and decides how to proceed. An autonomous robot that is *intelligent* and deserving of trust must be capable of being situationally aware; otherwise its decisions and actions will be questionable and prone to failure.

Value Judgment: The ability of an autonomous system to possess and act upon positive human values, reflected in behavior that is legal, ethical, and moral, is central to the trustworthiness of the system. For example, the 4D/RCS autonomous intelligent control system architecture, described below, has always possessed a value judgment module that interacts with its world model to generate behavior that is constrained by its values. In addition to values that can reflect human legality, ethics, and morality, the value judgment module also has function/mission priorities, i.e., work values important for the robot to carry out its task or mission. For example, the value of the autonomous robot's employing cover and concealment during a tactical approach may become secondary to the value of timeliness in reaching its objective; or the value of preserving its own existence may be deemed secondary to the value of accomplishing a particular mission. It is important for the trust of humans that the behavior of an autonomous robot (or robotic vehicle) have worthy values to constrain its behavior (unlike Skynet or the Terminator).

Context Cognition: The ability of an autonomous system to know the context of its status and mission allows the system to complete its mission despite, for example, adversarial disruption causing the loss of communications with the command center. Context cognition provides the system with expertise based on experience in similar situations, or prior knowledge relevant to the current situation, such as knowledge of the terrain despite the loss of GPS due to jamming. Just as context cognition allows a human squad cut-off from its platoon to continue to carry out its mission successfully, it can similarly benefit a squad of autonomous combat vehicles. The ability of an autonomous system to continue a mission despite adversarial interference in an anti-access/area-denial environment would strongly engender trust in the system.

Self-Reflexive Meta Control System: Reflexivity generally refers to the examination of one's own beliefs, judgments and practices. The self-reflexive, meta-control system provides self-awareness to the autonomous system. It is aware of its own states, actions, motives, goals, and action production mechanisms. It can engender meta-trustworthiness, where an autonomous system has ability to understand what a user needs to trust it, and the ability to assess trustworthiness of the user. The self-reflexive, meta-control system, with its understanding of itself and user could also be transparent to the user, able to divulge a complete accounting of its behavior. Self-reflexivity does not assure the success of a system's missions, but its ability to explain can lead to rapid improvements in performance and trust in its autonomy.