Can a Robot Do a Trust Fall?
Absurdity as a Component of Human Intelligence and Embodiment
Paper type: Short Paper

Amy LaViers and Ilya Vidrin
Mechanical Science and Engineering Department and Department of Theatre
University of Illinois at Urbana-Champaign and Northeastern University
Urbana, IL 61801 USA and Boston, MA 02115 USA
alaviers@illinois.edu and i.vidrin@northeastern.edu

Abstract

Trust is often considered valuable in a broad range of relationships, from professional collaborations to personal partnerships. This article examines the possibility of trust in a robotic system. By posing the question "can a robot do a trust fall?", an investigation on the issues embedded in designing trusting systems is presented, using methods and perspectives from philosophy and engineering. Posing such a question helps us understand the physicality and embodiment of trust, as well as the limits and resources of robotics.

Introduction

There have been a number of recent moves in the philosophy of emerging media to distinguish between the concepts of "reckoning" and "judgment". The distinction holds that "reckoning" is a limited form of "calculative rationality", while judgment is linked to full-blooded intelligence "that is existentially committed to its own existence and to the integrity of the world as world" (Smith 2019). The distinction between reckoning and judgment is particularly useful for differentiating what systems are capable of, including their limitations and resources.

One overlapping aspect of judgment and reckoning seems to be a criteria of rationality. Yet humans can and do behave irrationally. In fact, acting irrationally is a natural part of human experience. Moreover, we are able to learn and grow from our own irrationality (Ashcroft, Childs, and Myers 2016). Likewise, seemingly integral to human understanding and experience is the phenomenon of absurdity. One example of absurd behavior is risk-taking. While taking "calculated risks" are a necessary component of commonplace actions such as signing automotive or home loans and investing in higher education, there are more free-form, aesthetic risks that are just as ubiquitous. Bungee jumping and riding roller-coasters are two examples of risky behaviors that are motivated by aesthetic ends.

When we undertake a risk, we may trust that everything will be fine. When riding a roller coaster, for example, an individual can fully believe that the ride is not actually risky, thus placing a certain limited trust in the machinery. We may also not care about the results of our risky endeavors, such as if an individual does not care if things turn out poorly (in which case it does not matter what they believe or trust). Yet, without a *genuine* feeling of risk, the roller coaster has little aesthetic value to the rider.

Returning to judgment and reckoning, people who engage in risky behavior need not weigh options in a deliberative, "rational" manner. They can (and do) engage in irrational behavior in order to pursue particular ends (e.g., ecstatic states) in the moment. It is perhaps even possible that individuals can occupy two seemingly opposite states simultaneously: it is raining, but I do not believe that it is raining. This statement is an example of Moore's paradox (Moore 1993). While it is not something that we *reasonably* assert, we argue that there is still some value in such claims. After all, risky behavior is not necessarily rational.

Certain risky experiences can be relationship-building when undertaken together. For example, in a "trust fall" exercise, where one falls backwards into the arms of another, we can experience a moment of wild abandon. Rather than promoting our own individual homeostasis and livelihood, the exercise can serve as a ritual of putting trust in another person. Without the risk of falling to the ground, the exercise does not work, as we will outline further in this paper.

Perhaps this activity is an example of exemplification (Elgin 2010), wherein certain qualities of the experience of the action are non-propositional. Aesthetic ends that are manifest through creative, non-deliberative tasks may also require exemplification to advance understanding (Elgin 2010). Machines, governed by circuitry that implement first-order Boolean logic, are in some sense behaviorally limited by this. Thus, here, we propose a question: **can a robot do a trust fall?** In this paper we will examine the conditions of executing a genuine trust fall, followed by an examination of why creating such a protocol for a robotic system fails. We conclude by posing questions around the potential insights of this investigation.

Defining a "Trust Fall"

There are many variations of the trust fall exercise. One typical version occurs within a dyad: while standing, one person, the "faller", closes their eyes and falls backwards, relying on being caught by a "spotter" before hitting the ground. The exercise is based on the assumption that the act of reliance, when successfully executed, builds trust *between* the two individuals. A trust fall involves some kind of agreement on both sides – the faller agrees to fall and the spotter

agrees to catch. The full force of this agreement obligates the faller to be vulnerable and obligates the spotter to be receptive. If the faller does not care about being injured or is not afraid of hitting the ground, then the exercise will not function effectively to build trust between partners.

There are two paradigmatic cases worth exploring 1) the spotter does not succeed in catching the faller, and 2) the faller stumbles, flinches, or abandons the fall (consciously or not). In the first case, the spotter fails to uphold the obligation of catching. This is the potential fear of the faller. Even if the spotter is prepared and ready, the faller must allow themselves to be vulnerable. This can be quite an absurd task, in which one agrees to be vulnerable at the expense of not being caught or even getting injured. In the second case, the faller fails to *deliberately* fall. That is, in stumbling, bending their legs, or twisting around at the last moment to face forward, the faller "falls cautiously" and exemplifies a lack of trust in the catcher.

There are further problems inherent in the physical action. The exercise can become too "practiced", wherein the faller is no longer being genuinely vulnerable. By extension, the exercise is no longer "absurd". To promote vulnerability, the spotter may change at which point they catch the faller, such that the faller does not know when they will be caught. But the faller may not care about being vulnerable, such that even if the catcher is deliberately catching, the faller is not fully realizing the exercise of deliberately falling. Feelings of trust may thus be unidirectional, in that the faller trusts the catcher but the catcher does not trust the faller or vice versa.

We have an assumption that we are *not* designed to fall. As such, to fall is to place oneself at risk – a seemingly absurd action. The catcher is sensitive to the risk of the faller, just as the faller is sensitive to being caught. Thus it is not possible to establish *genuine* trust through physical interaction if the trust is unidirectional. This can be better illustrated through an example of catching a falling plate. Sensing that a plate is falling, I lean forward to catch it. Success in catching may lead to a sense of accomplishment or even augmented self-worth, but the action has not served to build trust *between* me and the plate.

We recognize that a "true" (as defined above) trust fall is a *rarity*. The exercise is contingent on skill and orientation of each agent toward the other, including such as factors as mood and environment. Thus a "true" trust fall cannot be arbitrarily created by any two humans on any given day. This excludes (the unlikely scenario) of tripping into a trust fall (versus deliberately falling). Notably, the component of trust features prominently in the initiation of falling on the part of the faller. Ostensibly, trust can only be built if the faller is genuinely vulnerable, meaning they are afraid of falling and choose to fall anyway. To design a system in this way means accounting for the way in which a system goes outside of its own protocol.

Formulating a Policy (Machine)

Moore's paradox provides an example that we have argued demonstrates how absurdity can be key to human understanding and experience. This section provides an attempt to create a computational structure that encompasses the case wherein a system holds both the proposition and its counter in validity at the same moment. To format this attempt, we will use the general structure of a transition system, which can describe the architecture of software-controlled engineered systems. In general, we think of this transition system as creating a discretized *policy* for our machine.

To model the possible structure for the action of the faller (noting that the catcher has a similar structure that we will not explicate here), let us define our transition system as a tuple of a set of 1) Q, states, 2) q_0, initial states, 3) E, events or "actions" that can be taken from a given state, 4) o, an output function that associates state-event pairs to a new system state, 5) Π higher-level propositions that are true or false at each state, and 6) h, a labeling function that associates states to propositions.

This definition will be used to describe the machine's current *state* and which states it may evolve to from that state, via which available actions. It provides a substrate onto which we can apply first-order predicate logic structure, enacting a series of propositions based on the structure of a formula ϕ, expressed in linear temporal logic. Specifically, we consider temporal operators *next*, $\mathbf{X}$, and *until*, $\mathcal{U}$, and logical operators *and*, $\wedge$, and *not*, $\neg$. Other logical operators like *or* $\vee$ and *implication* $\rightarrow$ can be expressed through combinations of these operators, i.e., disjunction is $\phi_1 \vee \phi_2 := \neg(\neg\phi_1 \wedge \neg\phi_2)$. Similarly, aggregated temporal concepts like *eventually* and *always* $\mathbf{G}$ can be defined, i.e., $\mathbf{G}\phi := \neg\mathbf{F}\neg\phi$. The formula $\mathbf{G}\alpha$ states that proposition α holds at all states of operation. Such a formula can be represented as a *Büchi Automaton* as in (LaViers et al. 2011). Changes in state and event structure can also be enacted to try and capture the phenomenon described in the previous section. We will use this abstraction to concretize our comments about trust through motion.

To create an initial model that may capture this desired behavior, assume that we have a sensing system in place that is able to detect gross physical states, i.e., by calculating and integrating the various measures of the machine sensors. These states are defined relative to the machine's stability, e.g., with a static stability detector, we may be considering simply whether the center of mass is inside the polygon of support or not. Similarly, on our event structure, we have an actuation system that computes steps needed to either *increase* or *decrease* stability[1].

Thus, the initial model of our system becomes:

$$T_1 = (Q_1, q_{1_0}, E_1, \Pi_1, h_1)$$

where
1) $Q_1 = \{\text{stable}, \text{unstable}\}$;
2) $q_{1_0} = \{\text{stable}\}$;
3) $E_1 = \{\text{stabilize}, \text{mobilize}\}$ offers two action options;
4) $o : Q_1 \times E_1 \mapsto Q_1$, as shown in Fig. 1;
5) $\Pi_1 = \{\text{falling}, \text{not falling}\}$; and

[1]The concern of this paper is not on creating a state-of-the-art stability detector or generator, so we leave these details to the imagination of the reader.

6) $h_1 : Q_1 \mapsto 2^{\Pi_1}$ makes associations between stability and falling, e.g., $h(\{\text{stable}\}) = \{\text{not falling}\}$ and $h(\{\text{unstable}\}) = \{\text{falling}\}$.

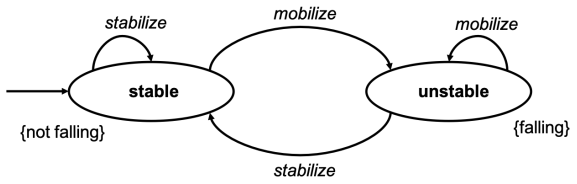

Figure 1: Visualization of T_1, showing state, event, and proposition structure.

We could create a robot that simply always mobilizes in an unstable state – this is a robot that would always fall over. Instead, we aim to design a robot that associates the instability inherent to the state of falling with both propositions "falling" and "not falling". We can use propositional logic to construct a controller for the system, creating a supervisory machine that will enforce the structure of these propositions as in (LaViers et al. 2011). For example, during a trust fall we could run **G**falling and when not engaged in that activity we could run **G**¬falling .

This abstraction of a machine will "function" correctly – it will stabilize itself when in an unstable state and "fall" in a trust fall. However, in a "trust fall", as discussed in the prior section, we want to create a state that can be pursued which is simultaneously "falling" and "not falling". Thus, an additional proposition needs to be associated with the action of mobilization. In this case, the faller needs to mobilize with a *true* vulnerability that falling may occur and, at the same time, a *true* confidence that it will be caught.

Thus, an updated model of our system might become:

$$T_2 = (Q_1, q_{1_0}, E_1, \Pi_1, h_2)$$

where

$h_2 : Q_1 \mapsto 2^{\Pi_1}$ makes associations between stability and falling with the added nuance that in some unstable situations the machine may also be "trusting", e.g., "not falling" as shown in Fig. 2.

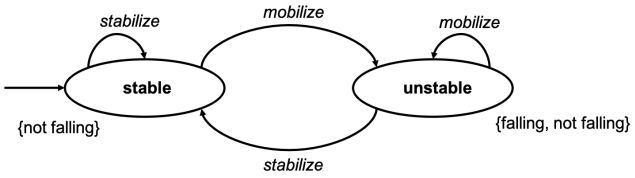

Figure 2: Visualization of T_2, showing state, event, and proposition structure.

Now, we need a controller that executes the proposition **G**(falling $\wedge$ not falling) (during a trust fall) and **G**¬falling (during a non-trust fall). We can see this fails because either the stable or unstable state satisfy "not falling", resutling in a robot that may always fall. Another attempt may create distinct states and events associated with this activity, e.g.,

$$T_3 = (Q_3, q_{1_0}, E_3, \Pi_3, h_3)$$

where $Q_3\{\text{stable}, \text{unstable}, \text{unstable but safe}\}$;

$E_3 = \{\text{stabilize}, \text{mobilize}, \text{mobilize cautiously}\}$

$\Pi_3 = \{\text{falling}, \text{not falling}, \text{falling and not falling}\}$

$h_3 : Q_3 \mapsto 2^{\Pi_3}$ makes associations between stability and falling with the added nuance that distinguishes unstable situations as shown in Fig. 3.

This structure would allow for a state designated as "falling and not falling" but it requires two distinct mobilization activities (to avoid being nondeterministic) and *a new unstable state*, meaning the machine is not in an *authentic* unstable state during the trust fall. Moreover, we must create a third proposition to label this state, as giving it both "falling" and "not falling" would result in the same error we encountered with T_2.

This example demonstrates a puzzling phenomenon for robotics and AI researchers to reconcile: curiosity seems to be an inherent and critical feature of natural intelligence. Yet, the act of being curious involves making mistakes, taking actions for "no reason", and, more generally, engaging in and being attracted to "play", activity that is inherently orthogonal to some other activity that may be defined as "work". (We could, also, build a robot designed to, say, play with children, then the "work" of this robot is, indeed, play.) Thus, the characteristic of being curious involves taking actions that are inherently off policy, and there is a circularity that we seem to not be able to get out of: if we create a policy that guides the machine to go off policy, the machine is simply obeying a broader policy from which it cannot deviate. Thus, can a machine enter the state of absurdity that is necessary to complete a genuine trust fall? Can a machine build genuine trust with humans, who seemingly do take on these states of vulnerability, if it cannot inhabit an off-policy position?

Discussion

Full-blooded trust requires some kind of judgment. Since the medium of the fall is movement, however, the judgment need not necessarily be deliberative. That is, the faller intentionally falls, without needing to *deliberate* by weighing options about when or how to do so. We understand a deliberative judgment as one that requires a serial process of weighing options, in which either individual (or both) assess the risks involved of falling and catching. A non-deliberative judgment is exercised in the experiential receptivity to the moment of falling. In other words, non-deliberative judgment is keyed into the true state. This need not be considered outside of conscious awareness (whatever that might mean biologically). That is, a feature of what we take to be *embodied reasoning* may or may not be something we are consciously thinking about, but it is still a feature of intelligence we are interested in examining.

We note that absurdity, curiosity, and vulnerability seem to be features of human intelligence that are keyed into embodiment, and that these features are required for activities like building trust and being creative. The trust fall is an

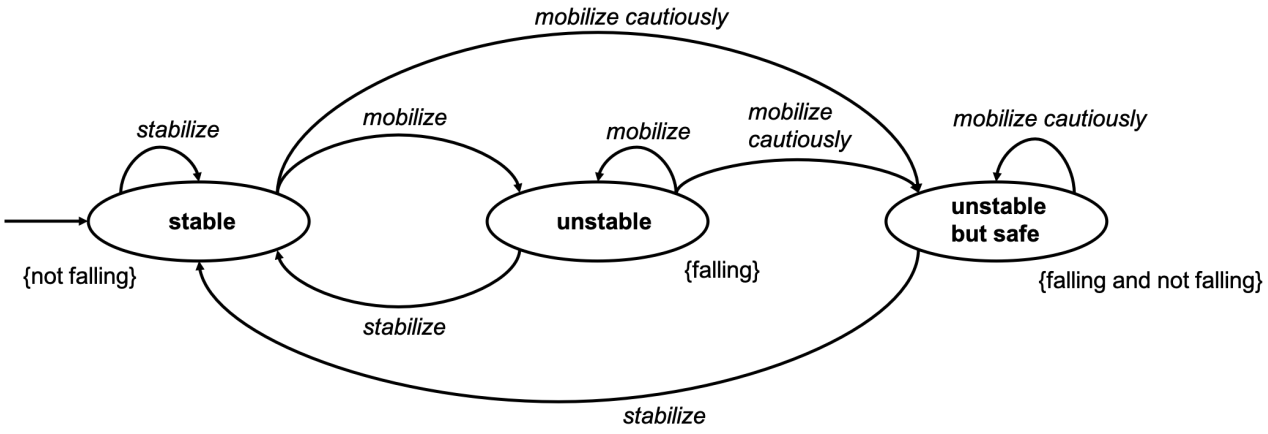

Figure 3: Visualization of T_3, showing state, event, and proposition structure.

exercise used in many community building and conflict mediation activities, but it is also a metaphor for many kinds of human relationship building activities, e.g., exchanging vows, sharing personal secrets, or appreciating ironic situations (e.g., rain on your wedding day). The formulation suggested here of a trust fall as an exercise requiring a moment of absurdity, vulnerability, and uncertainty from both agents ("Am I going to be caught?" and "Will I catch them?") and of a machine's policy as providing a forever level of reasonableness, resilience, and certainty, draw into question whether a machine can ever exhibit key features of human intelligence.

The trust fall seems to fall outside of first-order linear predicate logic. Specifying different qualities, textures, or even expectations means that trust is not just a product of a binary yes or no. Perhaps then, a trust fall establishes a bounded, three-part predicate of trust, such that the faller can reasonably say: I trust you *to catch me if I fall*. Even in successful circumstances in which the catcher has demonstrated their ability to catch the faller, it may be unreasonable for the faller to then assume complete trust in the catcher such that they say something like I trust you *to save me from a burning building*. Three-part predicates could also be further complicated by expressive qualifiers: I trust you to catch me *gracefully* or *carefully*. Exploring different logical structures may prove more fruitful to explain a genuine trust fall.

Conclusion

Modeling is a useful tool for gaining understanding of limitations, as much if not more than beneficial resources. The limits of modeling trust are located in the fact that trust has affective, expressive dimensions, but affective dimensions are not deterministic of mental attitudes. We argue that the affective dimension of trust makes it essential to *experience* rather than merely to *measure* it. Trust is not subject to a bounded binary of 'yes or no', but rather involves a richer set of behaviors. This means that accomplishing trust with computerized machines needs to account for the dynamic density of human experience, affording *genuine*, rather than simulated, reduced, or approximated, trust.

We argue that occupying absurd states helps us process ourselves and understand each other (the process of curiosity: truly going off policy and being uncertain). We have presented models of a trust fall in which p and $\neg p$ is internally consistent and suggest that perhaps exhibiting this paradox is an important piece of intelligence. This may highlight something humans do in embodied form, which may be termed somatic intelligence, emotional intelligence, intrinsic motivation, etc.

The notion that absurdity – or this internal consistency that breaks logic – may be necessary to the formation of *genuine trust* has implications for human-machine dyads or teams – especially in the growing field of human-robot interaction where machines are treated as agents inside social contracts. If machines cannot replicate the state of falling and not falling, then perhaps any notion of "trust" in these relationships requires new structures, terminology, and paradigms to be accurately described.

Acknowledgments

The authors contributed equally to the writing of this manuscript.

References

Ashcroft, J.; Childs, R.; and Myers, A. 2016. *The relational lens: Understanding, managing and measuring stakeholder relationships*. Cambridge University Press.

Elgin, C. Z. 2010. Exemplification and the dance. In ed. Roger Pouivet. Rennes., ed., *Philosophie de la Dance*. Presses Universitaire de Rennes.

LaViers, A.; Chen, Y.; Belta, C.; and Egerstedt, M. 2011. Automatic sequencing of ballet poses. *IEEE robotics & automation magazine* 18(3):87–95.

Moore, G. E. 1993. Moore's paradox. *GE Moore: Selected Writings* 207:212.

Smith, B. C. 2019. *The promise of artificial intelligence: reckoning and judgment*. MIT Press.