It Takes Two to Trust: Mediating Human-AI Trust for Resilience and Reliability

1st Juliette Zerick  
Intelligent Systems Engineering  
Indiana University  
Bloomington, Indiana, United States  
jzerick@iu.edu

2nd Zachary Kaufman  
Intelligent Systems Engineering  
Indiana University  
Bloomington, Indiana, United States  
kaufmazc@iu.edu

3rd Jonathan Ott  
Intelligent Systems Engineering  
Indiana University  
Bloomington, Indiana, United States  
jonott@iu.edu

4th Janki Kuber  
Intelligent Systems Engineering  
Indiana University  
Bloomington, Indiana, United States  
kuberjanki@gmail.com

5th Ember Chow  
Electrical & Computer Engineering  
University of Washington  
Seattle, Washington, United States  
emmychow@uw.edu

6th Shyama Shah  
Intelligent Systems Engineering  
Indiana University  
Bloomington, Indiana, United States  
shyashah@iu.edu

7th Dr. Gregory Lewis  
Intelligent Systems Engineering  
Kinsey Institute  
Indiana University  
Bloomington, Indiana, United States  
lewigr@iu.edu

Abstract—New technologies are destined to disrupt, but artificial intelligence (AI) has achieved unusual cultural impact, inspiring visceral fear in some, yet rapidly proliferating through pervasive adoption. But by its nature, adoption of AI necessitates more than mere acceptance: it requires trust. Trust surpasses cooperation; cooperation elicits predictability, but cannot enable the vulnerability described by human trust theorist Niklas Luhmann. The bond of trust must be developed through interaction, and humans inherit social constructs and societal norms that regulate verbal and behavioral communication, indicating internal state. AI must learn to read these cues to engender trust and recover when its human partner becomes distrustful. Our experimental platform, Hapti Bird, creates an environment for testing scenarios and observing interactions from which to better understand how humans trust each other, and how they trust AI. For this paper, we performed a variation of the Iterated Prisoner’s Dilemma with haptic devices and sensory capture involving two human subjects in an embodied joint action paradigm, taking the form of a video game. From video-derived heart rate (HR) and heart rate variability (HRV) we predict in-game cooperativity between subjects up to 7 seconds into the future (71% F1 score). Facial expressions tell of significantly different experiences of subjects depending on the amount of time given to establish trust, and when that trust is broken. Our accumulated findings educated an AI, named Hapti Bot, to embody the formulation of trust between humans. Hapti Bot was trained using genetic algorithms, which continuously generated AI candidates within given parameters. This process mimicked biological evolution and produced a Bot optimized to thrive in the Hapti Bird environment.

Index Terms—Artificial intelligence, human trust, human-machine interaction, psychophysiological markers, heart rate variability, polyvagal theory, genetic algorithms

The authors thank the Office of Naval Research and Crane Naval Surface Warfare Center for their support.

I. INTRODUCTION

As artificial intelligence (AI) rapidly evolves and proliferates to pervade daily life, fostering trust between humans and AI becomes paramount. Though the term broadly covers a wide array of use cases, we are particularly interested in intelligent machines that automate some human activity and are meant to interact with humans, or human-AI interactive systems (HAIS). A successful collaboration demands more than mere joint action, or coordinated behavior to bring about a change in the environment [1]. It requires trust, reflecting reliability not just of the AI, but of the human and the human-AI dyad, as well to enable a resilient connection.

Trust is a behavior seen as crucial to the functioning of human society [2]. It is a bond that allows humans to share vulnerability and take risks [3], enabling feats not otherwise achievable [4]. It is what underpins more complex social constructs such as culture, replete with knowledge and norms [5]—and how to be trustworthy, and identify others as such [3]. Biological evolution appears to have rewarded development of the cognitive capacity for such behaviors [6]. Even if the accumulation is not genetic, the resulting social structures we have today—government by consent of the governed [7], collective action, international courts—depend on trust. If trust is broken, bonds crumble and institutions fail. In an equivalent control system, if neither humans nor AI can maintain a trust state, collaborative activity ceases.

Guidelines for AI to be Trustworthy have been put forth at the highest levels [8], meant to establish a path for sustainable
development of AI. Common themes include human control, data privacy, accountability for the impact of AI systems, safety and security, transparency and explainability, fairness and nondiscrimination, professional responsibility, and promotion of human values in usage [9]. Further technical guidelines recommend reliability and resiliency [10], and a human rights-based approach advocates for equity in the reaping of benefits of advances in science and healthcare [11].

For AI to be Trusted in human society—in which it already functions, since AI drives trucks [12], recognizes faces for law enforcement [13], and provides companionship [14]—it must build and maintain trust with humans. Humans form relationships during interactions modulated by communication. While language is a powerful tool, not all communication is verbal and there is much information on a human’s internal state [15] that is communicated nonverbally via social cues, facial expressions, pose, gaze and gestures [16].

These are measurable psychophysiological markers. Facial action units (FAUs) are small expressions on the face, dozens of which are described and enumerated in the Facial Action Coding System (FACS) [17], and used extensively for classification tasks. Gaze tracking, sweating estimated with galvanic skin response (GSR), pose, and affect are further vectors of human-detectable social cues used in psychophysiological studies. Underlying the latter, more subtle cues is the human heart, the activity of which provides a window into both sympathetic and parasympathetic nervous systems [18]. As it is connected via vasculature to the entire body, modulation of the brain-heart network reflects low-level control of internal state; and observations of its activity contain rich information and memory. The body remembers even if the mind does not, and responds to events even if subjects are not consciously aware [19]. Thus heart rate (HR) and variance of times between beats, heart rate variability (HRV), are objective measures of the human internal state.

Equally important is detecting when humans stop trusting AI. Negative experiences with AI can result from unfulfilled promises of capabilities [20], displacement of employment, and feeling a loss of control, amongst other reasons. But there are consequences for seeming betrayal, and rebuilding trust requires work.

For an AI to recognize these psychophysiological markers and respond in kind to a human partner, an interface is required to mediate interactions. This communication channel provides a unique opportunity to satisfy some requirements of Trustworthy AI. With detection of markers correlated with trust (and distrust) as a measure of AI adherence to normative behavior, should the AI deviate from expectation, a human reaction can indicate a correction or recalibration is required to maintain requisite predictability, giving the system a degree of reliability [10]. Under the watch of a human partner who can identify the occurrence of unexpected adverse events, a sequence of translated human reactions provides a path through AI learning loss landscapes back to a stable dynamic. Such possible recovery provides the human-AI system resilience [10].

A. Experimental Studies in Human Trust and Human-Machine Interaction

Trust is more than just cooperation; it demands a refined comprehension of the multifaceted elements that underlie human behavior. In order to establish trust in AI systems, it is imperative to draw insights from the complex nature of interhuman trust. A comprehensive meta-analysis, incorporating over 2,000 studies was conducted to systematically evaluate the factors influencing human-human trust (primarily in teams and organizations), concluding that the reputation of the trustee and the shared closeness between truster and trustee as pivotal predictors of trustworthiness [21]. Here, we see that reliability and relationships build trust.

In instances of human-AI interactions where a high level of trust was demonstrated, the benefits of such trust have highlighted the potential for mutual reliance between humans and AI systems. In a design experiment where an AI agent was developed to manage the design process, track progress, and bridge communication in multidisciplinary teams, results indicated that teams under AI management demonstrated performance equal to or even superior to those managed by humans [22]. Team members perceived the AI agent as equally sensitive to the team’s needs as a human manager, suggesting its potential to match human capabilities in trustworthy management.

Understanding the dynamics of trust in AI is crucial as humans increasingly engage with AI systems in new contexts. When considering cooperation and competition during joint ventures with AI, studies have revealed that despite efforts to “humanize” AI, uncertainty persists in human-AI collaboration, which can negatively impact trust [23]. However, the cooperative nature of gameplay has shown to increase trust between a human player and an AI [22]. These findings demonstrate the complex relationship between human trust and the dynamics of cooperation and competition in the evolving landscape of human-AI interactions—and the clear need for an objective measurement platform with which to understand them. Hapti Bird is an initial effort at establishing just such a tool.

B. The Prisoner's Dilemma and the Iterated (or Repeated) Prisoner’s Dilemma

Our experiment is a variation of the Prisoner’s Dilemma, a thought experiment originating from game theory [24], where two prisoners await interrogation, but are held separately and unable to communicate with each other. Each has received an offer of leniency in exchange for testimony about the other. So the prisoners face the following dilemma: either remain silent or testify against their partner. If one prisoner betrays a silent partner, the prisoner will go free, while the partner receives a long prison sentence. If both offer to testify, both receive moderate sentences. If both say nothing, each receives a light sentence. Barring sentiment or further encounters, there is a clear incentive for betrayal.
The Iterated (or Repeated) Prisoner’s Dilemma (IPD or RPD, respectively) follows a similar structure; however, in this case, the two prisoners undergo the experiment repeatedly. It has been shown that incentives change when the prisoners repeatedly face this dilemma and literature in game theory has provided a plethora of strategies for this scenario [25].

C. Genetic Algorithms

The power of genetic algorithms (GAs) for finding optimal solutions is Darwinian selection applied to computational problems, not biological organisms. Inspired by nature’s selection process, GAs efficiently tackle diverse optimization problems with flexible rules [26] [27].

At their heart, GAs are dynamical systems [28] operating in a space of sets of parameters—a parameter space—to be optimized for some end. Trajectories of points in this parameter space are propelled by fitness functions and perturbed by mutation, but naturally fixate on attractors in the fitness landscape, should they exist.

The algorithm begins with a population of individuals, each of which has an encoded chromosome, which experiences evolutionary pressures from a fitness function determining individuals’ utility, or proximity to a solution. The fittest individuals reproduce to create new generations, so adaptation is a necessary condition for survival. After many generations, the population should have migrated to one or more possible solutions [26]. Due to the potential for mutation, genetic algorithms have an uncanny ability to rapidly find solutions, even in large, high-dimensional spaces.

D. The Heart and its Role in Cooperation and Competition

The human brain regulates the heart through the autonomic nervous system via sympathetic and parasympathetic branches [29]. The heart maintains a constant internal state of homeostatic balance between numerous demands. When there is a need to engage actively with the environment, cortical neurons shift this homeostatic balance, and cardiac output is increased to match metabolic demands [30]; the heart beats faster. Moreover, brain-heart bidirectional co-regulation can be altered by central autonomic commands, including those associated with stress, physical activity, arousal, sleep, and altered states of consciousness [31]. When a calm behavioral state is required, the re-engagement of cranial nerve X slows the heart rate and provides the physiological support for self-soothing behaviors [30] [32].

Heart rate variability (HRV) is derived from the heart rate, and a reliable tool to detect the activity of autonomic nervous regulation. Control and coordination of the body and brain is well reflected in the measurements of HRV [33] [34]. Low HRV has a direct correlation with increased stress and anxiety, which can hinder relationships. Contrarily, high HRV is an indication of greater stress resilience and promotes more positive social interactions, and is associated with better emotional regulation and empathy, all of which could facilitate trust and cooperation [34].
Fig. 1. Three stills of the rendered game during an experiment depict possible states. In a), the two players, with distinct color-coded cursors, approach an obstacle from left to right, tethered via simulated spring. b) shows the change in color of the obstacle (now green) to indicate the players have successfully navigated through one of the two available gates offering safe passage through the obstacle, earning resources. c) shows the alternative scenario, whereupon players have instead crashed into the obstacle (now red), losing resources.

which translates streamed video frames into raw pulse data by estimating blood flow in the face [35]. Produced data are highly sensitive to changes in lighting and subject movement. Raw data requires extensive cleansing and segmentation, with segments further reduced as both HR and HRV are slow measures, requiring at least several seconds of uninterrupted data to compute, and shifting only incrementally with each heartbeat.

Another camera simply records the subject. From this data we can identify FAUs. There are many FAU detectors available, but after some testing, we selected OpenFace [36] for its portability and reliability; there are newer and more accurate tools available but they did not meet other criteria for use in our study. Such detectors are classifiers that input images, or a sequence of frames from a video and determine whether any of some subset of FAUs are present. From our video data, we produced a stream of detected FAUs and their intensity (on a scale of 1-5). This data was nearly continuous but sparse; FAUs are comparatively rare events. Thus our metrics are typically frequencies of observed FAUs during a prescribed window of time.

C. Hapti Bot

The ultimate goal of this work is to capture human psychophysiological response to an AI counterpart by first understanding human-human trust dynamics. The AI counterpart we named Hapti Bot. While AI agents in games often employ reinforcement learning [37], the complexity of interaction in the game prevented the algorithm from converging to an optimal solution. GAs wrapped around a small neural network model proved more fruitful for developing reliably performing AI agents. The internal neural network required only two layers to intelligently select next moves. The model is given the game’s current state, which includes the individual’s position and orientation toward the oncoming obstacle, and the individual’s previous move.

The Hapti Bot lifecycle is depicted in Figure 2. Weights for the internal model became an individual’s chromosome. As the internal model trained, its fitness gradually peaked. When all individuals reach this performance plateau, the fittest are selected for producing the next generation. New, random individuals—which are AI candidates within given parameters—are also spawned and added to the population. This process continues until an individual emerges with a performance similar to humans, capable of passing through 30 gates without crashing into an obstacle.

IV. RESULTS

43 experimental runs were completed, with as participants one Subject, typically a university undergraduate student taking courses in psychology, and one laboratory volunteer, referred to as the Confederate, with knowledge of the game.

A. Heart Rate and Heart Rate Variability

The HR and HRV baselines were computed from a segment extracted prior to the start of the game. The ratio of HR and HRV to baseline proved to be the most effective data reduction approach for these models. Only 10 experimental runs contained sufficient data, but from them, an unusual model was constructed: from a subject’s immediate HR, HRV, and its comparison to baseline, we could predict their cooperativity, as expressed in cursor distance, up to 7 seconds in the future, with an F1 score of 71%. The inclusion of additional FAU measures (as depicted in Figure 3) improved the F1 score by a single point.

B. Facial Action Units

For this analysis, we split subjects by whether their experimental run began with an aligned (12 subjects) or unaligned
phase (17 subjects). If a subject experienced a phase of alignment, that could provide time to build trust, a buffer against the first taste of betrayal of an unaligned phase. Some subjects did not have this period, and after the introduction immediately jumped into an unaligned phase. For each subject, we observed the change in expression (frequencies of FAUs) during the transition to the unaligned phase by taking the difference between FAU frequencies 20 seconds before and after the transition. With a paired differences test per FAU, we identified FAUs that changed significantly during the transition, as shown in Figure 4.

We did not find significantly expressed FAUs within subgroups or the entire group of subjects, perhaps due to the small sample size, but more likely due to sheer diversity. Some subjects during experiments became hostile, even pounding the desk, while others appeared to truly enjoy the chaos of unaligned phases.

C. Hapti Bot

Currently, our implementation of Hapti Bot is able to locate the best gate and pass through it without collision with the obstacle, and repeat that success at least 1000 times, well above what is asked of any human participant. Hapti Bot is also able to correctly position itself for passage through the gate in approximately 60-70% of occurrences. After passage, the Bot immediately seeks out the next one, so that it will have substantial flexibility adjusting to gap length variation between gates and human player movements.

V. DISCUSSION

With measurable psychophysiological markers, we can observe the building and breaking of human trust in two contexts:
Fig. 4. In this table we compare the change in expression of FAUs of two groups of subjects during the transition from an introductory (I) or aligned (A) phase to an unaligned (U) phase; it is at this time that the subjects would experience betrayal. The first group experienced an introductory phase, an aligned phase, and then an unaligned phase. The second group did not have the benefit of an intermediate aligned phase, leaving little time for subjects to develop trust.

1. **Facial Action Unit (AU) Number**

<table>
<thead>
<tr>
<th>Facial Action Unit (AU)</th>
<th>Number</th>
<th>I → A → U</th>
<th>I → U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inner Brow Raiser</td>
<td>1</td>
<td>None</td>
<td>Increase</td>
</tr>
<tr>
<td>Outer Brow Raiser</td>
<td>2</td>
<td>None</td>
<td>Increase</td>
</tr>
<tr>
<td>Brow Lowerer</td>
<td>4</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Upper Lid Raiser</td>
<td>5</td>
<td>Increase</td>
<td>None</td>
</tr>
<tr>
<td>Cheek Raiser</td>
<td>6</td>
<td>None</td>
<td>Decrease</td>
</tr>
<tr>
<td>Lip Tightener</td>
<td>/</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Nose Wrinkler</td>
<td>9</td>
<td>None</td>
<td>Decrease</td>
</tr>
<tr>
<td>Upper Lip Raiser</td>
<td>10</td>
<td>None</td>
<td>Decrease</td>
</tr>
<tr>
<td>Lip Corner Puller</td>
<td>12</td>
<td>None</td>
<td>Decrease</td>
</tr>
<tr>
<td>Dimpler</td>
<td>14</td>
<td>None</td>
<td>Decrease</td>
</tr>
<tr>
<td>Lip Corner Depressor</td>
<td>15</td>
<td>None</td>
<td>Increase</td>
</tr>
<tr>
<td>Chin Raiser</td>
<td>17</td>
<td>None</td>
<td>Decrease</td>
</tr>
<tr>
<td>Lip stretcher</td>
<td>20</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Lip Tightener</td>
<td>23</td>
<td>Increase</td>
<td>Increase</td>
</tr>
<tr>
<td>Lips part</td>
<td>25</td>
<td>Increase</td>
<td>Decrease</td>
</tr>
<tr>
<td>Jaw Drop</td>
<td>26</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Lip Suck</td>
<td>28</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Blink</td>
<td>45</td>
<td>None</td>
<td>Increase</td>
</tr>
</tbody>
</table>

For the group that begins in an unaligned phase, we see smiles (AU codes 6, 12) wiped from faces in a clear sign of betrayal. Also significantly observed were instances of blinking (AU 45), an autonomic reaction associated with a startled or defensive response.

### C. Hapti Bot

The minimal and versatile nature of Hapti Bot allows it to be a point of inception for future versions of the Hapti Bird game. Its current form, in which every action it takes favors the best gate, will reasonably pass as human to real human subjects. However, there is much we can build on to explore more human-like traits.

### VI. Future Work

The Falcon controllers, while ideally precise and responsive, were discovered to be quite fragile in the hands of competitive university students. For the next version of Hapti Bot, we are using robust, rugged steering wheels as game controllers. We will validate the first round of experiments before integrating Hapti Bot for training. We expect its behavior to adjust, as it will now find itself utilizing a wheel for movement as well as an elastic tether to a human player.

Further, we plan to create two trained Hapti Bots, connected by game controllers, with opposing incentives for gate selection. By creating a small penalty for selecting a suboptimal gate, this new round of training may provide the Bot is the ability to "give up" and choose to pass through the suboptimal gate when it is no longer possible to reach its preferred gate, lest risking crashing into the obstacle.

With emerging Hapti Bots and further confirmation of psychophysiological markers, the logical next step is to create human-in-the-loop (HITL) machine learning systems [39]. Such systems iteratively integrate human feedback to improve models. In our case, Hapti Bot engages a human player who is observed in real-time for identifiable markers; such reactions are used to adjust Hapti Bot’s behavior via online reinforcement learning [40]. Unsupervised, Hapti Bot would become an autonomous cyber-physical system (CPS), an altogether different creature, reflecting the individual proclivities of its user [41].

### ACKNOWLEDGMENT

This work was funded by the U.S. Department of Defense (Contract W52P1J12093009). The authors thank the Office of Naval Research and Crane Naval Surface Warfare Center for their support.

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute.

### REFERENCES
