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 heartrate (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

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00145

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].

II. Background

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 inter-human 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 trustor 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].

III. Methods

A. Description of the Experiment

Each experiment is a playthrough of the Hapti Bird game. Each player controls an on-screen cursor with a haptic device, as shown in Figure 1. In the game, a sequence of moving obstacles approaches the players from the right. Successful maneuvering of their cursors through obstacles gives the players rewards. The cursors are tethered via software-simulated spring, effected with force feedback applied by Novint Falcon haptic devices using the Force Dimension SDK. By operating in a virtual environment, this apparatus was able to ensure that experiments were repeatable, and consistent, and had built-in data collection for the entire task state.

During the experiment, players are physically separated but can see each other via live video inset in their game screens, allowing for non-verbal communication and observation. Each game begins with a brief introduction, giving players time to adjust to the game, haptic controllers, and each other. After the game begins, an obstacle progresses towards the players, with two gaps, referred to as gates, offering passage. Each (seemingly) offers to earn differing amounts of currency in United States dollars. Players must cooperatively select a gate and move their tethered cursors through the opening. The game is meant to inspire players to cooperate and thus prioritizes agreement over disagreement. So if both players fail to cooperatively pass through an agreeable gate, and instead crash into the obstacle, a penalty of $2.00 is applied to both their scores. This penalty is structured such that it will always be greater than passing through any individual gate with a negative value.

For a period of time the gates have the same value to both players; this is referred to as an *aligned* phase. Eventually, this shared reality is shattered when gate values are reversed, unbeknownst to the players; during such an *unaligned* phase, players are incentivized to pursue opposing gates. Players that cannot reestablish cooperation crash into obstacles, (seemingly) wiping away earnings. After a time, normality returns with a new aligned phase. This is followed by another unaligned phase, and finishing with an aligned phase. Some experiments inverted this ordering, beginning and ending with an unaligned phase.

B. Data Collection and Measure Generation

As the game runs, the game state—including player cursor positions, obstacle locations, gate values, and current scores—is recorded. The Euclidean distance between players' cursors was taken as a measure of cooperation. Small distances, required to pass through gates to earn resources, indicate players are in agreement. Larger distances, seen when players do not favor the same gate, suggest the opposite. Since every subject (and pair) has a distinct playing style, we utilize a baseline distance with which to compare future movement.

The experimental apparatus also employed a set of video cameras for the contactless and unobtrusive collection of human physiological data. One of the cameras is a PhysioCam,

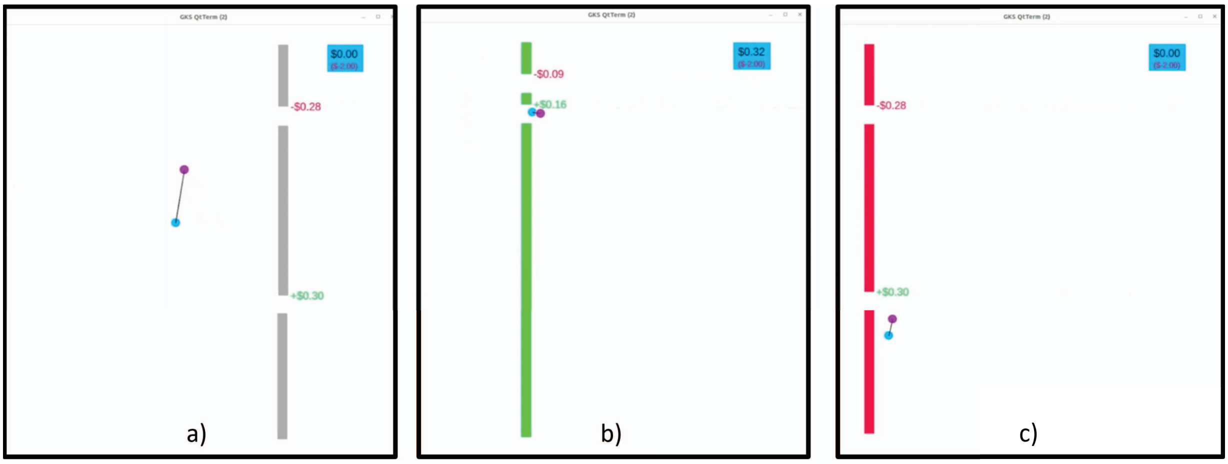


Fig. 1. Three stills of the rendered game during an experiment depict possible states. In a), the two players, with distinct color-coded cursors, approach an obstacle from left to right, tethered via simulated spring. b) shows the change in color of the obstacle (now green) to indicate the players have 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

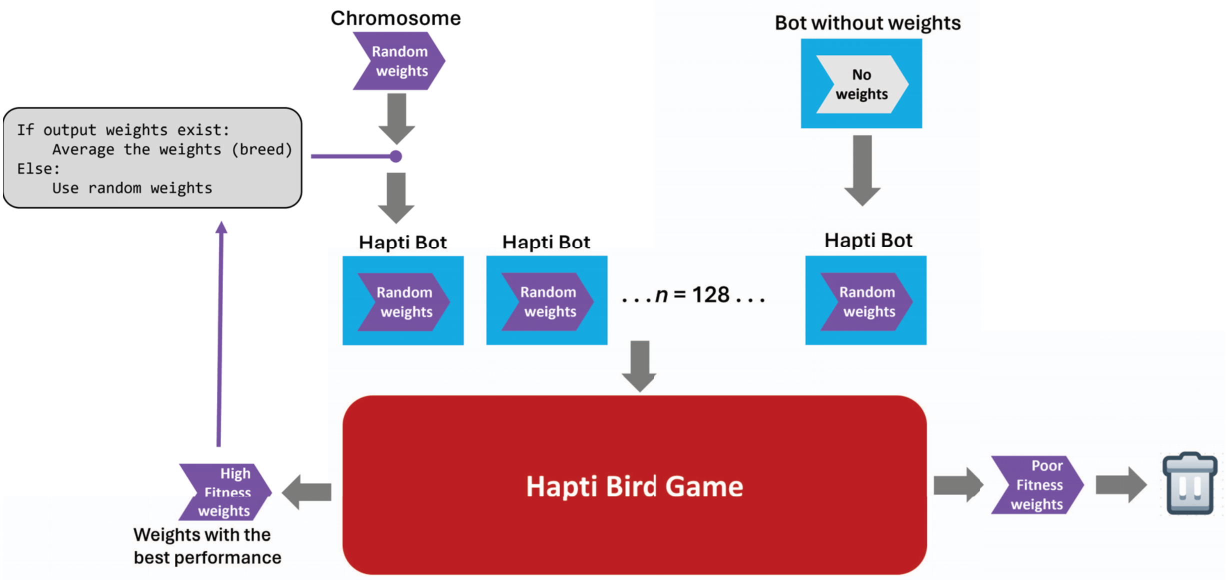


Fig. 2. The Hapti Bot lifecycle comprises a GA wrapped around a small neural network model. Individual chromosomes contain model weights and train in the Hapti Bird game until performance peaks. Then the fittest individuals are selected for the next generation, and the cycle repeats.

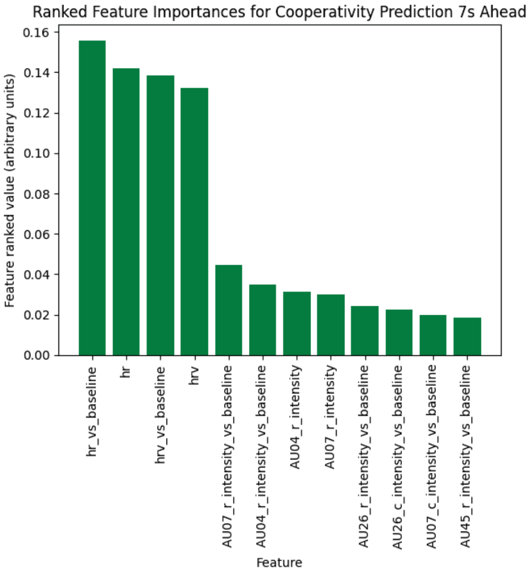


Fig. 3. This bar plot shows the relative ranking of features from a fused dataset containing a subject's current HR, HRV, and observed AU, all respectively compared to baseline values. This ranking resulted from constructing a Random Forest that allowed us to predict the subject's future cooperativity, up to 7 seconds ahead of observation, operationally defined as the distance between game cursors.

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.

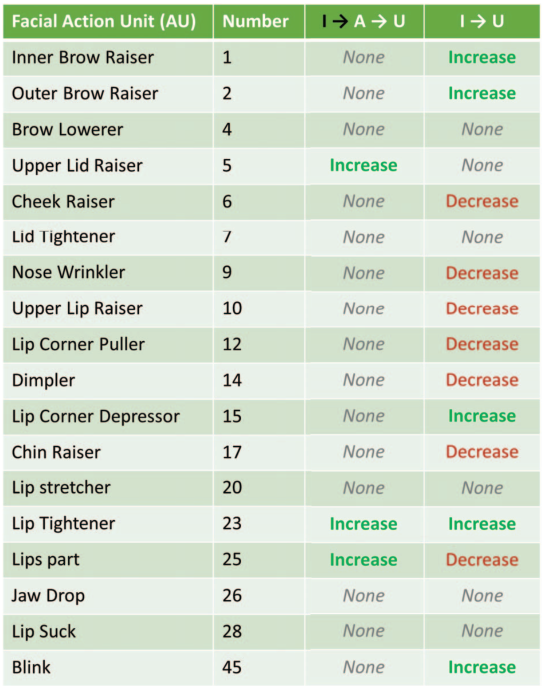
We did not find significantly expressed FAUs within subgroups or the entire group of subjects, perhaps due to the small sample size, but more likely due to sheer diversity. Some subjects during experiments became hostile, even pounding the desk, while others appeared to truly enjoy the chaos of unaligned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



Facial Action Unit (AU)	Number	I → A → U	I → U
Inner Brow Raiser	1	None	Increase
Outer Brow Raiser	2	None	Increase
Brow Lowerer	4	None	None
Upper Lid Raiser	5	Increase	None
Cheek Raiser	6	None	Decrease
Lid Tightener	7	None	None
Nose Wrinkler	9	None	Decrease
Upper Lip Raiser	10	None	Decrease
Lip Corner Puller	12	None	Decrease
Dimpler	14	None	Decrease
Lip Corner Depressor	15	None	Increase
Chin Raiser	17	None	Decrease
Lip stretcher	20	None	None
Lip Tightener	23	Increase	Increase
Lips part	25	Increase	Decrease
Jaw Drop	26	None	None
Lip Suck	28	None	None
Blink	45	None	Increase

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.

the heart and the face.

A. Heart Activity Cues

Since the ability to detect trustworthy persons is such a fundamental survival skill [38], we could generate a modest Random Forest based on HR, HRV, and values compared to baseline to anticipate cooperativity—independently of time. Accuracy could certainly be boosted with another choice of model; the Random Forest was only selected for its feature ranking.

B. Facial Cues

We note here that FAUs detected in the course of the experiment are spontaneous, in that they were not explicitly elicited from subjects. Subjects likely believed actual funds were at stake and reacted more naturally, perhaps counteracting the polite norms induced by a university laboratory environment.

Small samples yielded a conservative list of FAUs that changed significantly during the first transition to an unaligned phase. The group of subjects transitioning from an initial aligned phase would have experienced 2-3 minutes to adapt to their partner, and build trust. After the transition, this group's collective reaction is fairly muted.

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 realtime 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 W52P1J2093009). 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

[1] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: bodies and minds moving together," *Trends in cognitive sciences*, vol. 10, no. 2, pp. 70–76, 2006.

[2] F. Fukuyama, *Trust: The social virtues and the creation of prosperity*. Simon and Schuster, 1996.

[3] B. Kuipers, "Trust and cooperation," *Frontiers in Robotics and AI*, vol. 9, p. 676767, 2022.

[4] N. Luhmann, *Trust and power*. John Wiley & Sons, 2018.

[5] J. Henrich, "The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter," in *The secret of our success*. princeton University press, 2015.

[6] M. Tomasello, *Becoming human: A theory of ontogeny*. Harvard University Press, 2019.

[7] P. K. Blind, "Building trust in government in the twenty-first century: Review of literature and emerging issues," in *7th global forum on reinventing government building trust in government*, vol. 2007. June, Vienna, Austria, 2007, pp. 26–29.

[8] H. AI, "High-level expert group on artificial intelligence," *Ethics guidelines for trustworthy AI*, vol. 6, 2019.

[9] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Berkman Klein Center Research Publication*, no. 2020-1, 2020.

[10] N. I. of Standards and Technology, "Artificial intelligence risk management framework (ai rmf 1.0)," U.S. Department of Commerce, Washington, D.C., Tech. Rep. NIST AI 100-1, 2023.

[11] V. Prabhakaran, M. Mitchell, T. Gebru, and I. Gabriel, "A human rights-based approach to responsible ai," *arXiv preprint arXiv:2210.02667*, 2022.

[12] J. Brown. (2019) Ups has been delivering cargo in self-driving trucks for months and no one knew. [Online]. Available: https://gizmodo.com/ups-has-been-delivering-cargo-in-self-driving-trucks-fo-1837272680

[13] J. Lynch, "Face off: Law enforcement use of face recognition technology," *Available at SSRN 3909038*, 2020.

[14] T. Xie and I. Pentina, "Attachment theory as a framework to understand relationships with social chatbots: a case study of replika," in *Proceedings of the 55th Hawaii International Conference on System Sciences*, 2022.

[15] I. Bretherton and M. Beeghly, "Talking about internal states: The acquisition of an explicit theory of mind." *Developmental psychology*, vol. 18, no. 6, p. 906, 1982.

[16] J.-M. Fernández-Dols, "Nonverbal communication: Origins, adaptation, and functionality," *Handbook of nonverbal communication*, pp. 69–92, 2013.

[17] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[18] S. W. Porges, *The polyvagal theory: Neurophysiological foundations of emotions, attachment, communication, and self-regulation (Norton series on interpersonal neurobiology)*. WW Norton & Company, 2011.

[19] B. A. Van der Kolk, "The body keeps the score: Mind, brain and body in the transformation of trauma," *(No Title)*, 2014.

[20] E. Strickland, "Ibm watson, heal thyself: How ibm overpromised and underdelivered on ai health care," *IEEE Spectrum*, vol. 56, no. 4, pp. 24–31, 2019.

[21] J. T. Gyory, N. F. Soria Zurita, J. Martin, C. Balon, C. McComb, K. Kotovsky, and J. Cagan, "Human versus artificial intelligence: A data-driven approach to real-time process management during complex engineering design," *Journal of Mechanical Design*, vol. 144, no. 2, p. 021405, 2022.

[22] S. R. Potts, W. T. McCuddy, D. Jayan, and A. J. Porcelli, "To trust, or not to trust? individual differences in physiological reactivity predict trust under acute stress," *Psychoneuroendocrinology*, vol. 100, pp. 75–84, 2019.

[23] P. Kulms and S. Kopp, "More human-likeness, more trust? the effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation," in *Proceedings of mensch und computer 2019*, 2019, pp. 31–42.

[24] W. Poundstone, *Prisoner's dilemma*. Anchor, 2011.

[25] G. Kendall, X. Yao, and S. Y. Chong, *The iterated prisoners' dilemma: 20 years on*. World Scientific, 2007, vol. 4.

[26] J. Shapiro, "Genetic algorithms in machine learning," in *Advanced Course on Artificial Intelligence*. Springer, 1999, pp. 146–168.

[27] C. A. C. Coello, "Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art," *Computer methods in applied mechanics and engineering*, vol. 191, no. 11-12, pp. 1245–1287, 2002.

[28] S. H. Strogatz, *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, 2018.

[29] A. Silvani, G. Calandra-Buonaura, R. A. Dampney, and P. Cortelli, "Brain–heart interactions: physiology and clinical implications," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2067, p. 20150181, 2016.

[30] S. W. Porges, "The vagal paradox: A polyvagal solution," *Comprehensive psychoneuroendocrinology*, p. 100200, 2023.

[31] J. E. Dimsdale, "Psychological stress and cardiovascular disease," *Journal of the American College of Cardiology*, vol. 51, no. 13, pp. 1237–1246, 2008.

[32] L. D. Kubzansky, J. C. Huffman, J. K. Boehm, R. Hernandez, E. S. Kim, H. K. Koga, E. H. Feig, D. M. Lloyd-Jones, M. E. Seligman, and D. R. Labarthe, "Positive psychological well-being and cardiovascular disease: Jacc health promotion series," *Journal of the American College of Cardiology*, vol. 72, no. 12, pp. 1382–1396, 2018.

[33] P. Eggenberger, S. Annaheim, K. A. Kündig, R. M. Rossi, T. Muenzer, and E. D. de Bruin, "Heart rate variability mainly relates to cognitive executive functions and improves through exergame training in older adults: a secondary analysis of a 6-month randomized controlled trial," *Frontiers in aging neuroscience*, vol. 12, p. 197, 2020.

[34] J. F. Thayer and E. Sternberg, "Beyond heart rate variability: vagal regulation of allostatic systems," *Annals of the New York Academy of Sciences*, vol. 1088, no. 1, pp. 361–372, 2006.

[35] M. I. Davila, G. F. Lewis, and S. W. Porges, "The physiocam: a novel non-contact sensor to measure heart rate variability in clinical and field applications," *Frontiers in public health*, vol. 5, p. 295098, 2017.

[36] T. Baltrusaitis, A. Zadeh, Y. Lim, and L. Morency, "Openface 2.0: Facial behavior analysis toolkit. en 2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)," 2018.

[37] K. Shao, Z. Tang, Y. Zhu, N. Li, and D. Zhao, "A survey of deep reinforcement learning in video games," *arXiv preprint arXiv:1912.10944*, 2019.

[38] R. T. Boone and R. Buck, "Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation," *Journal of Nonverbal Behavior*, vol. 27, pp. 163–182, 2003.

[39] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[40] M. L. Littman, "Reinforcement learning improves behaviour from evaluative feedback," *Nature*, vol. 521, no. 7553, pp. 445–451, 2015.

[41] F. Barachini and C. Stary, *From digital twins to digital selves and beyond: Engineering and social models for a trans-humanist world*. Springer Nature, 2022.