ARL-TR-9628 • JAN 2023

DEVCOM ARMY RESEARCH LABORATORY

Long-Term Analysis of a Human-AI Collaboration Study Using a Mobile Game

Torin Adamson, Arvind Pillai, Andrew Campbell, Lydia Tapia, Lidia S Obregon, and Evan C Carter

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-9628 ● JAN 2023

Long-Term Analysis of a Human-AI Collaboration Study Using a Mobile Game

Torin Adamson and Lydia Tapia
University of New Mexico

Arvind Pillai and Andrew Campbell
Dartmouth College

Lidia S Obregon
University of Arizona

Evan C Carter
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE		Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
January 2023	Technical Report	1 October 2021 – 30 September 2022

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Long-Term Analysis of a Human-AI Collaboration Study Using a Mobile Game	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)	5d. PROJECT NUMBER
Torin Adamson, Arvind Pillai, Andrew Campbell, Lydia Tapia, Lidia S Obregon, and Evan C Carter	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FA Aberdeen Proving Ground, MD 21005	ARL-TR-9628

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES

ORCID ID: Evan C Carter, 0000-0001-7471-8769

14. ABSTRACT

Human-AI interaction is typically studied in laboratory settings where participants spend minutes to hours per session. The benefits of this approach are well-known: excellent experimental control and sophisticated sensing of behavioral and non-behavioral data (e.g., psychophysiology) to name a few. The downsides are discussed less frequently. The burden on experimenters and participants from sessions, difficulties in collecting massive data sets, and the inability to analyze long-time-scale processes limit the produced knowledge. Our experimental setup combines mobile devices and wearable sensors as a means of addressing these limitations. Via a mobile game application called "Busy Beeway," for studying human-AI collaboration, we require participants to balance their strengths and weaknesses with autonomous partners in obstacle avoidance tasks. Participants provide game data daily over several months while their "context"— heart-rate, activity, sleep, environment—is measured continuously by StudentLife. In this technical report, we present an initial survey of interaction styles that participants developed over time with different autonomous capabilities and across changing context. Additionally, we present how this data can be used to predict human-AI collaborative performance, thus providing a way to suggest human-AI pairings.

15. SUBJECT TERMS

Human-Autonomy Teaming, Mobile Sensing, Adaptive Autonomy, Mobile Videogames, Ambulatory Assessment, Humans in Complex Systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Evan C Carter
Unclassified	Unclassified	Unclassified	UU	27	19b. TELEPHONE NUMBER (Include area code) (240) 478-9295

Contents

List of Figures

List of Tables

1. Introduction

Existing human-automation motion planning studies in small, controlled laboratory environments yield valuable knowledge[1–3]; however, to the best of our knowledge, larger studies that run over long periods of time are absent from the literature. As a companion to the high-fidelity research data obtained in conventional laboratory studies, this "in the wild" approach would improve human-automation systems in ways previously unreachable. Through the use of technology widely ubiquitous to the average consumer, such as smart phones and wearable computers, participation in such long-term, mobile studies becomes feasible. There have already been efforts to adapt human-automation study environments into a mobile game[4] and to develop a passive continuous data collection platform employing consumer-grade wearable sensors.[5] Systems like these provide the necessary building blocks with which to design and deploy large-scale, human-automation studies.

This technical report presents preliminary results in an ongoing large-sample, longitudinal study where human participants play a dynamic obstacle avoidance game while swapping control with an autonomous agent. Participants sign up and install "Busy Beeway," the mobile game component of the experimental setup, and wear Garmin sensors connected to the "StudentLife" application to provide additional data that reflects the context around the lives of each player. The evaluation presented here is primarily focused on the result of assigning differently configured AI partners on game play performance, to discover any emergent patterns of behavior, and to determine if the biological context of the player can impact their game play. We find a wide variety of game play patterns that are dependant on the particular player, the assigned AI, and sometimes even a change in a player's biological context. This suggests a need to find methods in future work that can predict what AI configuration produces the most desired results based on the combination of all these factors per individual.

2. Background

Shared control systems allow humans and automation to combine their strengths while covering each others' weaknesses[6]; however, it is necessary to validate these systems with human subject studies. Existing laboratory studies can provide insight into the effect of trust in human behavior with automation,[2,7] ways of non-verbal communication between the human and agent,[3,8–10] and how the human operator

can react to degradation or failure of automation.[1,11,12] While these small laboratory environments allow for precise control over experimental variables, there exists a large domain of knowledge that cannot be reached under these conditions, particularly how behavior can change over time with repeated exposure to the system or from external factors in participants' lives.

Attempting to perform laboratory studies on massive numbers of participants or with many repeated measurements over time is infeasible. With video games now a part of our culture,[13] they can provide a creative solution to this problem where human-automation studies could be adapted into video games, also known as "gamification." Previous work has explored the potential of supplementing conventional laboratory studies with games based on human-automation collaboration tasks, using dynamic obstacle avoidance as an example.[4] Gamification introduces its own set of problems, however, as the system must provide rewards that are either intrinsic or extrinsic and encourage continued participation.[14] Existing video games that already provide entertainment for humans while also involving human-AI teams would be excellent environments for future studies.[15] This work uses an obstacle-avoidance mobile game specifically designed for long-term daily participation with experimental parameters that can be updated live as needed.

Collecting data from participants "in the wild" as they go about their daily lives sacrifices the tight experimental control of the typical laboratory study. This potential increase in unknown effects on participants can be a benefit to the extent that enough data are collected and the correct covariates are also captured. To this end, a mobile device in tandem with a wearable sensor can provide basic biometrics along with the administration of self-report surveys.[5] Our work augments Busy Beeway[4] with data provided from StudentLife[5] with additional self-reporting surveys to measure attention and participation compliance.

3. Methods

3.1 Study Protocol

We sourced the data for this preliminary work from an ongoing human subjects study involving Busy Beeway and StudentLife. So far, 58 participants were recruited from a convenience sample of students and graduate students at the University of Arizona. After an initial screening, participants were invited to the lab for an onboarding session during which the components of the data collection system were explained and loaded onto their phones. Participants also completed a battery of self-report items (not analyzed here) and a brief training built into Busy Beeway that was designed to ensure an understanding of the game. Other than specific troubleshooting required for some participants, this was the only in-person interaction participants had with members of the study team.

Participants were prompted to provide data once per day for 180 days. The prompt occurred at either 7 AM, 12 PM, or 5 PM local time, and participants had 3 h to open the game application. Participants were not able to play the game outside of these time windows. At the start of a daily session, the participant first answered several self-report questions designed to assess "state variables," such as affect and sleep. Participants then played three levels of Busy Beeway, each with progressing difficulty. At the end of three levels, participants could upload their data. Daily sessions were designed to take no more than 5 min on average.

We designed a monetary incentive scheme to motivate participants to provide data throughout the long time scale of our data collection. Compensation was tied to daily uploads, specifically, participants were paid a base rate of 0.25 USD for the first 30 uploads with a 0.25 USD increase every 30 uploads (i.e., 0.25 USD for uploads 1–30, 0.50 USD for uploads 31–60, 0.75 USD for uploads 61–90). Furthermore, we provided participants with a 3.00 USD bonus if they uploaded data for 3 days in a row. All compensation was provided through peer-to-peer payment applications (e.g., Venmo). Compensation was administered as soon as possible by the study team, typically within several days of uploading data.

To motivate participants to engage with the game throughout a session, we also designed a bonus that was meant to reward attention. One of six in-game items were randomly placed near the second or third sub-goal in each level. Upon completion

of all levels, participants were quizzed on which items they had seen. This bonus quiz indirectly motivates a degree of performance, as well, in that a certain level of success is required to get to the point in the level where every bonus item was visible.

3.2 Busy Beeway

Busy Beeway is a mobile research video game designed for the study of human-automation collaboration for motion planning.[4] The game involves navigation and collision-avoidance in the face of stochastic, moving obstacles. The version used in this study allows for remote collection of game play data and survey responses, and experimental parameters can be changed and delivered to active participants through the app as needed. Participants for this study were given instructions during an onboarding process along with the link to download and install the app via Google Play as a closed beta.

In Busy Beeway, the participant navigates a bee avatar (1 in Fig. 1) from flower to flower (2) and then to a portal (3) to complete the level. (We refer to the flowers as sub-goals and the portal as the final goal.) The participant is told to complete the task as quickly as possible and they are shown a timer (4) throughout game play. In addition to speed, the participant must also focus on safety, since there are stochastically moving obstacles (5), represented as wasps. Upon collision between the avatar and a wasp, the participant loses a "life" (7) and must start the level again. Each level can be attempted four times. The participant controls the avatar through a touch-screen-based joystick (6), and upon removal of their finger, the avatar is controlled by an autonomous driving assistant. An arrow (8) near the avatar points in the direction of the nearest goal at all times.

3.3 Data Collection

Our data collection follows a highly repeated within-subjects design. The primary manipulation of interest are the AI "personalities" that were assigned on each day. We implemented AI "personalities" as four parameterizations of artificial potential fields (APFs),[16] designed to vary across their ability to complete a level safely versus quickly. Each personality is color-coded in the game (the orange outline of the avatar in Fig. 1) and participants are told upon recruitment that the different colors correspond to different AIs. The parameters for each of these AI agents are shown in Table 1.

Fig. 1 Game play interface of Busy Beeway. (1) Bee avatar, (2) sub-goals represented as flowers, (3) final goal represented as a portal, (4) current play time, (5) dynamic obstacles represented as wasps, (6) on screen joystick, (7) current state of control and the number of re-tries or "lives," (8) arrow pointing to the nearest goal.

Table 1 Gaussian APF parameters assigned to each of the four AI configurations used in this study. "Sigma" is the standard deviation using in-game distance units and the "Repulsion" is the factor that prioritizes evading obstacles over traveling toward the goal.

AI	Color	Sigma	Repulsion
1	Orange	0.55	72%
2	Light blue	0.62	98%
3	Dark blue	1.40	98%
4	Purple	1.00	75%

The game application records all raw screen input and all information about the game objects (position, heading, etc.). Data are recorded at 30 Hz, which matches the per-second frame rate the game maintains. In addition to game play, participants are also given a self-report questionnaire daily. An example of how these items were implemented is given in Fig. 2 and the questions are listed in Fig. 3. This survey asks for the time the participant went to sleep and woke up, the quality of their sleep, and the degree to which they feel certain emotions on a Likert scale.

(a) Analog Scale

(b) Likert Scale

(c) Time

Fig. 2 Examples of the different types of questionnaire prompts in Busy Beeway. Analog scale questions are from 0.0 to 1.0, five-point scale (Likert) questions are multiple choice, and time questions ask for a specific hour and minute of day.

1. About what time did you go to sleep?
2. About what time did you wake up?
3. How would you rate your sleep quality?
4. How was your mood when you woke up?
5. How alert did you feel when you woke up?
6. How upset do you feel?
7. How interested do you feel?
8. How determined do you feel?
9. How irritable do you feel?

Fig. 3 The daily survey given to participants. Items 1 and 2 required a time input, items 3 through 5 presented an analog slider response, and items 6 through 9 were on a five-point Likert scale.

In addition to the data collected through the game application, participants were also measured using a wearable sensor (Garmin vivoactive 4) and a suite of passive sensing software uploaded to their devices. The sensing software is known as StudentLife[5,17] and is used to collect data via a mobile phone's onboard sensors (e.g., GPS, microphone). In this study, StudentLife was modified to also coordinate the collection of data from the Garmin, which includes a variety of sensors: an accelerometer, a barometric altimeter, a compass, GPS, a gyroscope, a heart-rate monitor, and a pulse oximeter. Here, we focus only on step counts taken from the Garmin and its estimate of raw heart rate. Please see our previous report[18] for additional details on the data we collect.

3.4 Data Analysis

Using data collected from subjects under the protocol in Section 3.3, we are able to analyze subject interaction with the AIs under several conditions. In this section, we introduce the metrics used for this analysis including those that reflect performance (e.g., collisions, speed) and those that characterize interaction (e.g., quantity and direction of interaction). The data provides both analysis of a subject's behavior along with a potential data source to be used as features for prediction of a subject's performance with a given AI.

3.4.1 Data Processing

The data collected in this study have varying units from step counts to heart rates in beats per minute (BPM). The data were standardized to range from 0.0 to 1.0 through normalization. Some data sources contain missing sections that occur if the participant fails to adhere to the study protocol (forgetting to put the wearable sensor on, etc.) or a malfunction in their mobile device occurs.

The daily data includes the ratio of manual control versus automated control, moods and sleep durations recorded via self-report (see Fig. 3), heart rate data in BPM, the count of steps taken, measured sleep duration from the Garmin device, and movements via GPS. The data from the control ratios and daily moods can be noisy and a locally weighted smoothing technique (LOESS) is used, specifically MATLAB's rloess method. A window width, w, must be defined. Unlike the rest of the data, the control ratios are already within range and do not get normalized. For heart rate data, the minimum, average, and maximum BPM are calculated for each day along with two Hurst exponents, α_1 α_2, extracted from detrended fluctuation

analysis (DFA).[19] The ranges are normalized separately from the exponents. To determine the amount of sleep each participant had each day, the duration of sleep is recorded both from the wearable Garmin device and from the daily survey (Fig. 3). The data from the Garmin is used if available, if not, the value self-reported by the participant is used. Due to the confusion of selecting the correct value for AM or PM in the survey, strange outliers in self-reported data may exist; therefore, negative sleep durations or extreme sleep durations (above 16 h) were excluded. To create a one-dimensional measure based on the participant's location as tracked by GPS, the geometric center of all the data collected is calculated and the distance from this center is used.

The metric of performance used in this study is the relative collision rate of the human-AI system against the expected collision rate of the AI acting alone. The values for the AI-only collision rates are shown in Table 2. The relative collision rates for a particular AI is subtracted from the mean collision rate, resulting in positive values indicating worsening collision rates when compared to AI alone with negative values indicating a reduction in collision rates. This data is then smoothed with LOESS using the same window value, w.

3.4.2 Observations

To create the set of observations, the collected data throughout each participant's study period is paired with the collision rate data and discretized into multiple observations that fit within a sliding window. The length of this window, w, matches the length used to smooth the data. In each window, linear regression is performed on the data from the past w days and the slope of the trend is recorded in x_n for all $n = 16$ sources of data. The average value for collision rates within the window is used for y_j for all $j = 4$ AI configurations. This forms a single observation. The window is adjusted 1 day at a time until the entire study period is covered with overlapping windows, resulting in $180 - w$ observations for a 180-day study period per person.

Table 2 Statistics on the performance of each AI agent alone per session in terms of collisions occurred and time for successful attempts. Levels were collected by simulating AI agents alone for 120 sessions (30 per each agent). The standard deviations are listed next to each value. Later levels contain more obstacles.

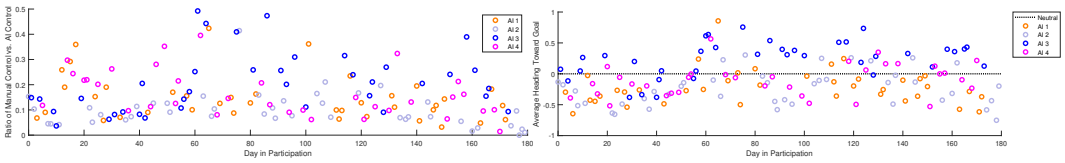
AI	Collisions	Completion Time (s)		
	Overall	Level 1	Level 2	Level 3
1	6.77 ± 2.38	20.2 ± 0.60	20.5 ± 0.35	20.1 ± 0.70
2	1.83 ± 1.46	21.6 ± 1.16	23.3 ± 1.26	24.7 ± 2.04
3	1.97 ± 1.59	26.0 ± 2.73	33.1 ± 3.92	38.3 ± 5.71
4	9.84 ± 2.02	19.8 ± 0.72	20.1 ± 0.30	19.7 ± 0.32

4. Results and Discussion

To understand evolving player behavior over time and to find feasible methods for making human-AI pairing decisions, the game play and biological context data gathered thus far are examined. First, a single player is used as an example for performing behavioral data analysis. Then, data from the 21 players who have completed the study are used to predict changes in collision rates for human-AI pairs.

4.1 Analyzing Player Behavior

Player 12's changing game play behavior over time readily demonstrates the importance of long-term studies. The basic characteristics of player behavior can be recorded through their decision to allow the AI agent to continue or to intervene and assert manual control. In Fig. 4a their ratio of asserted control with each AI agent is illustrated over the study period. While player 12 maintains occasionally high intervention ratios for AIs 1 and 4, it is around the 60-day mark where a dramatic shift in asserted control with AI 3 occurs. They fluctuated between 3.7% to 20.5% each session before this point, and between 9.1% to 49.2% afterwards. This action is not beneficial to the AI agent as the collision rate of the resulting human-automation system exceeds what would be expected from AI 3 alone by one standard deviation as seen in Fig. 5. Player 12 maintains this behavior despite experiencing 118 collisions after this change when AI 3 would be expected to only encounter 45 in the same time frame. This motivates two questions: why did Player 12 change their strategy around the 60-day mark and why did they choose to assert more control over AI 3?



(a) Interventions (b) Heading

Fig. 4 Ratio of manual control asserted by player 12 over the course of the 180-day study period (a) along with average heading towards the next goal at the end of player 12's input events (b). Each dot represents a single daily session. A heading is calculated as the dot product between the player's normal movement direction and a normal pointing toward the next goal. A value of 1 means to head directly towards the goal and –1 means to head directly away.

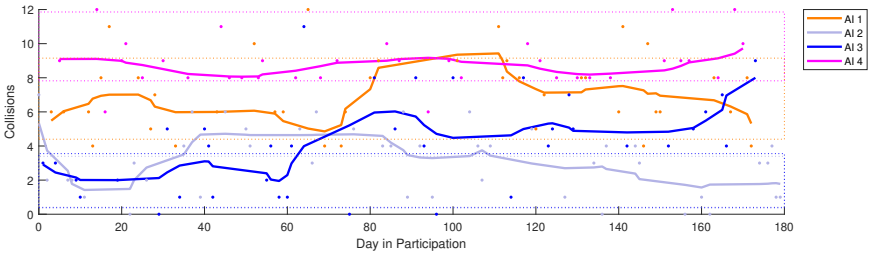


Fig. 5 Collisions experienced each session group by AI. The dots represent the raw data, the trends are generated with a 14-day LOESS smoothing window, and the shaded rectangles indicate the amount of collisions expected from each AI alone. The dotted rectangular region indicates one standard deviation in either direction.

We hypothesize that player 12 began to prioritize level completion over obstacle avoidance for two reasons. First, AI 3 takes the most time among the other configurations to avoid obstacles. Second, the headings the player directs the controlled avatar to are biased towards the next goal. Fig. 4b displays the average heading at the end of every input event per session over the course of the study. A normal is calculated for the direction the player avatar moves in and its dot product is taken against a normal pointing toward the next goal. Around the 60-day mark, we see headings at the end of interventions pointing the avatar more often toward the next goal for AI 3 (in blue). The average heading does not drop below -0.02 and indicates such a prioritization for level completion.

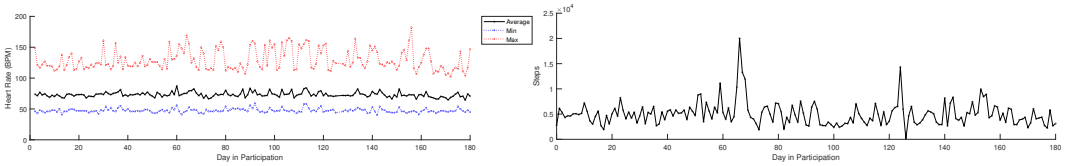
External factors in the daily life of players could have an influence on their behavior and the continuous sensor data gathered from StudentLife should be considered. For player 12, the heart rate and step counter data collected from the Garmin wearable are shown in Fig. 6. Past the 60-day mark, this participant attained daily maximum heart rates more often than in days before, but the step counter only indicates two

major periods of higher physical movement. Before the 60-day mark, the heart rate only exceeded 150 BPM on 3 days, and on 29 days after. This difference is highly disproportionate even considering the latter time frame taking up two thirds of the period. The amount of walking detected also reaches a maximum step count toward day 66, but returns back to a usual level–only momentarily rising on day 124.

Some event or change in the player's life might explain what is seen here. This highlights the importance of biological context as it could potentially have had an impact on their game play in addition to any learning and development of strategies from within the game; however, the nature of the effect and its strength must be explored in future study. Healthy people can have a higher heart rate during psychological stress,[20] and there can be a link between psychological elements (stress and emotions) and physiological activity (heart rate, heart rate variability, etc.)[21] Heart rate brought on by factors other than physical activity (non-metabolic heart rate) is an excellent indicator of one's emotional state.[22] As there was no increase in physical activity as indicated by step count, player 12's heart rate increase may be due to psychological reasons.

4.2 Predicting Human-AI Performance

Our goal is to develop a system that suggests human-AI pairings such that team performance is improved. We formulate this problem as one of supervised learning. In this report, we evaluate the accuracy of collision rate predictions for human-AI pairs based on the past 14 days of observations against the ground truth. To perform this, every observation taken from a player starting at day 14 (the number of days that must pass to begin recording observations), a nearest neighbor search using a root-mean-square deviation (RMSD) distance metric calculated on the trends, x, in the training set to find all observations within a radius, r. Because there may be missing periods of data (participation absence, device malfunction, etc.) not all 166 overlapping window observations may exist for some players. Different radii are used for different players, the ones selected for this study are listed in Table 3. To find these, first the radius is set to 0.02 and decremented by 0.001 until RMSD queries for that player fail (returning no results); the smallest successful radius is then chosen. The average values among observations returned by the query are used as the prediction of relative performance.



(a) Heart Rate (b) Step Counter

Fig. 6 Heart rate data collected from player 12 over the study period (a). DFA is shown above and average heart rate with ranges are shown on the bottom. The steps taken each day for this participant is shown in (b).

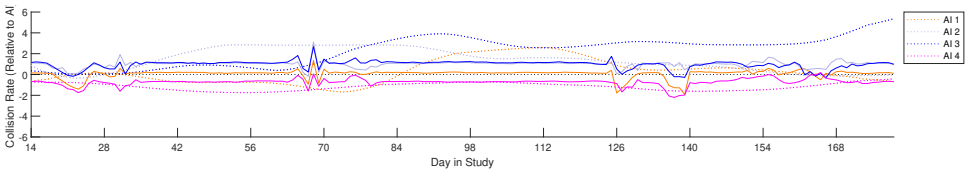
Table 3 Recorded data and collision rate performance observation pair counts, query radii, example prediction accuracy, and total amount of asserted manual control for each player to date in this study. Players will have fewer than 166 observations when data is missing (periods of participation absence, etc.). RMSD query radii are empirically chosen to be just large enough for queries to be possible (at least one result returned). Accuracy is measured in the number of days in which the recommended AI in terms of relative collision performance reflects the actual performance for that player vs. total observed days.

Player	Observations	Radius	Accuracy	Manual control
9	160	0.010	30.6%	2.94%
11	166	0.010	78.3%	10.40%
12	166	0.008	89.2%	15.31%
13	166	0.011	63.3%	1.58%
15	136	0.008	27.2%	10.71%
16	165	0.010	98.8%	14.31%
17	159	0.010	28.9%	3.09%
18	166	0.014	21.1%	2.64%
19	166	0.009	16.9%	5.09%
20	166	0.012	63.3%	9.56%
21	162	0.012	13.6%	6.67%
22	163	0.009	62.6%	8.80%
23	120	0.010	17.5%	1.16%
24	164	0.008	88.4%	30.19%
25	129	0.013	75.2%	35.31%
26	166	0.009	98.8%	11.07%
30	166	0.011	88.0%	40.20%
31	166	0.013	100.0%	44.05%
33	166	0.007	44.0%	7.37%
34	166	0.008	90.4%	3.33%
35	166	0.013	81.3%	21.43%

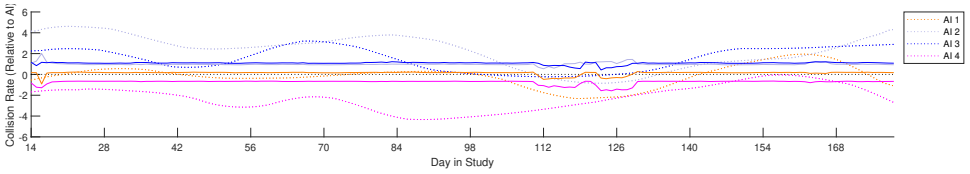
Fig. 7a illustrates an example of this test on player 12. If selecting for the AI configuration to partner with this player, the AI predicted to have the lowest (preferably negative) relative collision rate would be selected. The accuracy of this prediction can be measured by the number of days where the predicted lowest collision rate and the observed lowest collision rate suggest the same AI configuration versus to-

tal observation days. If using player 12 as an example, the prediction agrees with the observation 89.2% of the time. Most of the time, the prediction returns the same ordering of relative human-AI performances reflecting the common outcomes; AIs 2 and 3 are difficult to improve upon and AIs 1 and 4 are easier. As a more extreme example, it is difficult to predict the collision rates for player 21 as they never achieved significant positive or negative impacts (see Fig. 7c). Player 31 on the other hand achieved collision rates that match the average achieved by all players almost entirely consistently. At some point after the 154-day mark, the observed collision rates for AIs 1 and 2 switch places for player 31 (see Fig. 7b); however, the prediction is accurate as AI 4 still had the most improvement. Indeed, it is usually a matter of deciding between AI 1 or 4 for predictions in collision rate improvement when paired with players.

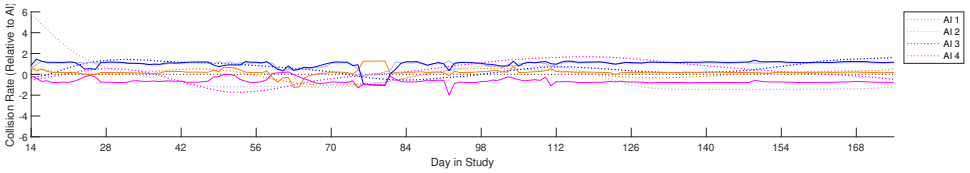
This prediction method not only relies on the continuous data provided by the mobile device and wearable sensor, but the game play behavior of the player over the past 14-day window. The player chooses when to take control from the AI and if a player does not provide much manual input, the collision performance will resemble the AI playing alone more closely. This causes the relative collision rates to tend toward zero, resembling the AI playing alone. Player 21 is a notable example with only having controlled the avatar 6.67% of the time. The relationship between prediction accuracy and total manual control is shown in Fig. 8. For lower amounts of manual control, the prediction task becomes difficult and results in variable degrees of accuracy. However, for larger amounts of control (about 15% or more), accuracy tends to be high. Thus, a minimum amount of engagement with the human-AI game is necessary if the continuous sensing context data is to be useful, otherwise it ends up being more closely related to the AI itself, which produces inconsistent results.



(a) Player 12



(b) Player 31



(c) Player 21

Fig. 7 Comparison of actual relative collision rates in a 14-day average window (dotted lines) with predicted average collision rates (solid lines) over the course of several players' study periods grouped by AI. Collision rates are relative to those each AI would experience alone. Player 12 (a) has a prediction accuracy of 89.2%. Player 31 (b) represents the most accurately predicted player (100%), while player 21 is the least (13.6%) (c).

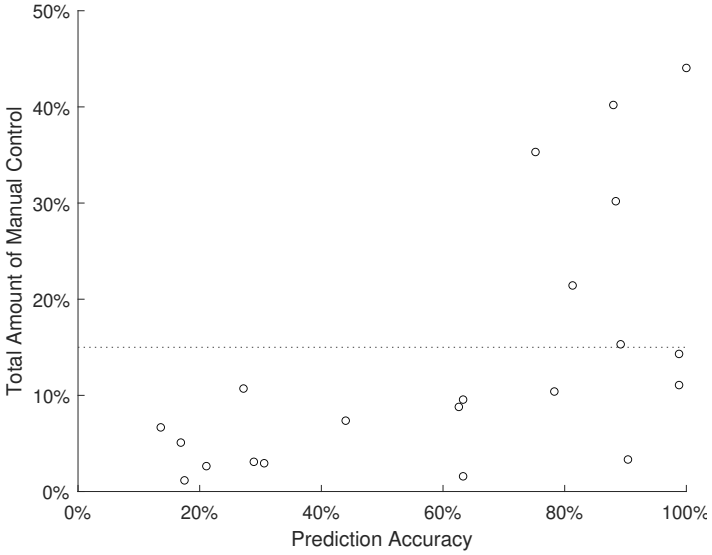


Fig. 8 The prediction accuracy of each player in this study to date vs. their total amount of manual control asserted throughout the study period. While an accurate prediction does not necessarily indicate high amounts of player control, those that have at least 15% (dotted line) tend to be easier to predict.

5. Conclusions

This ongoing large, longitudinal study has demonstrated that participant interactions with AI agents can change over time. In the example shown, we demonstrated that one participant increased interactions with a risk-adverse AI after almost 2 months within the study through interventions that steered the agent more directly toward goals. Additionally, the recorded biometrics demonstrated increased heart rate readings during the time period when participant behavior changed with no indication of increased physical activity. This, along with self-reported daily responses, suggests that the individual's psychological stress may have triggered the change in behavior toward the AI. This case study suggests a direction for future analyses of our entire sample, once collected.

Using aggregated data across all participants, we also demonstrated that prediction of participant performance with the AI agent, as given through a measure of expected amount of collision, can be done. A proof of concept using a nearest-neighbor approach demonstrates high prediction accuracy when a participant provides sufficient examples for interaction.

6. References

1. de Waard D, van der Hulst M, Hoedemaeker M, Brookhuis KA. Driver behavior in an emergency situation in the automated highway system. Trans Hum Fact. 1999;1(1):67–82.

2. Xu A, Dudek G. Maintaining efficient collaboration with trust-seeking robots. In: Proc IEEE Intl Conf Intel Rob Syst (IROS). 2016. p. 3312–3319.

3. Cutlip S, Wan Y, Sarter N, Gillespie RB. The effects of haptic feedback and transition type on transfer of control between drivers and vehicle automation. IEEE Trans Hum-Mach Sys. 2021;51(6):613–621.

4. Adamson T, Oishi M, Chiang HTL, Tapia L. Busy Beeway: a game for testing human-automation collaboration for navigation. In: ACM Proc Intl Conf Motion Games (MIG). 2017 Nov. Barcelona, Spain. 2017. p. 9:1–9:6.

5. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, Zhou X, Ben-Zeev D, Campbell AT. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: ACM Proc Intl Jt Conf Pervasive Ubiquitous Comput (UbiComp). 2014 Sep. Seattle, WA. 2014. p. 3–14.

6. de Winter JC, Dodou D. Preparing drivers for dangerous situations: a critical reflection on continuous shared control. In: IEEE Intl Conf Sys Man and Cyber. 2011. p. 1050–1056.

7. Yin W, Chai C, Zhou Z, Li C, Lu Y, Shi X. Effects of trust in human-automation shared control: a human-in-the-loop driving simulation study. In: IEEE Intl Int Transp Sys Conf. 2021. p. 1147–1154.

8. Johns M, Mok B, Sirkin DM, Gowda NM, Smith CA, Talamonti Jr WJ, Ju W. Exploring shared control in automated driving. In: ACM Intl Conf Hum Rob Interaction (HRI). 2016 Mar. Christchurch, New Zealand. 2016. p. 91–98.

9. Hudspeth M, Balali S, Grimm C, Sowell RT. Effects of interfaces on human-robot trust: specifying and visualizing physical zones. In: Proc IEEE Intl Conf Rob Auto (ICRA). 2022. p. 11265–11271.

10. Fernandez-Fernandez R, Aggravi M, Giordano PR, Victores JG, Pacchierotti C. Neural style transfer with twin-delayed DDPG for shared control of robotic manipulators. In: Proc IEEE Intl Conf Rob Auto (ICRA). 2022. p. 4073–4079.

11. Li R, Li Y, Li SE, Zhang C, Burdet E, Cheng B. Indirect shared control for cooperative driving between driver and automation in steer-by-wire vehicles. IEEE Trans Intl Transp Sys. 2021;22(12):7826–7836.

12. Huang C, Lv C, Hang P, Hu Z, Xing Y. Human–machine adaptive shared control for safe driving under automation degradation. IEEE Int Transport Sys Magazine. 2022;14(2):53–66.

13. Granic I, Lobel A, Engels RCME. The benefits of playing video games. American Psychologist. 2014;69(1):66–78.

14. Eveleigh A, Jennett C, Blandford A, Brohan P, Cox AL. Designing for dabblers and deterring drop-outs in citizen science. In: Proc SIGCHI Conf Hum Factors Comput Syst (CHI). 2014 Apr. Toronto, Ontario, Canada. 2014. p. 2985–2994.

15. Verhoeven Y, Preuss M. On the potential of rocket league for driving team AI development. In: IEEE Symp Series Comput Int. 2020. p. 2335–2342.

16. Khatib O. Real-time obstacle avoidance for manipulators and mobile robots. In: Proc IEEE Intl Conf Rob Auto (ICRA). Vol. 5. 1986 Mar. St Louis, MO. 1986. p. 90–98.

17. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, Zhou X, Ben-Zeev D, Campbell AT. StudentLife: using smartphones to assess mental health and academic performance of college students. In: Mobile Health. Springer; c2017. p. 7–33.

18. Adamson T, Wang W, Hasan Y, Campbell A, Tapia L, Carter E. A mobile data collection system for studying human autonomy teaming in conjunction with passive context and psychophysiological sensing. DEVCOM Army Research Laboratory (US); 2021. Report No.: ARL-TR-9359.

19. Peng CK, Havlin S, Stanley HE, Goldberger AL. Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos. 1995;5(1):82–7.

20. Lambiase MJ, Dorn J, Chernega NJ, McCarthy TF, Roemmich JN. Excess heart rate and systolic blood pressure during psychological stress in relation to metabolic demand in adolescents. Biol Psychol. 2012;(91):42–47.

21. Brown RS, Brosschott FJ, Versluis A, Thayer J, Verkuil B. New methods to optimally detect episodes of non-metabolic heart rate T variability reduction as an indicator of psychological stress in everyday life. Int J Psychophysiol. 2018;:30–36.

22. Brouwer AM, Dam vE, Erp vBJ, Spangler PD, Brooks RJ. Improving real-life estimates of emotion based on heart rate: a perspective on taking metabolic heart rate into account. Frontiers Hum Neuro. 2018;12(284):1–10.

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
APF	artificial potential field
app	application
BPM	beats per minute
DFA	detrended fluctuation analysis
GPS	global positioning system
LOESS	locally weighted smoothing
RMSD	root-mean-square deviation

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

1 DEVCOM ARL
(PDF) FCDD RLB CI
TECH LIB

1 DEVCOM ARL
(PDF) FCDD RLA FA
E CARTER