Expert Systems With Applications 199 (2022) 117167

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Detection of driver health condition by monitoring driving behavior through machine learning from observation

Avelino J. Gonzalez [a,*], Josiah M. Wong [a], Emily M. Thomas [a], Alec Kerrigan [a], Lauren Hastings [a], Andres Posadas [a], Kevin Negy [a], Annie S. Wu [a], Santiago Ontañon [b], Yi-Ching Lee [c], Flaura K. Winston [d]

[a] Computer Science Department, University of Central Florida, 4000 Central Florida Blvd., Orlando, FL 32816, USA
[b] Computer Science Department, Drexel University, Philadelphia, PA, USA
[c] Psychology Department, George Mason University, Fairfax, VA, USA
[d] Department of Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
Machine learning from observation
Modeling and simulation
Automotive safety
Driers afflicted with ADHD
Health monitoring
Artificial Intelligence

ABSTRACT

This paper describes our investigation to determine whether undesirable health conditions of an automobile driver can be identified in real time solely by monitoring and assessing his/her driving behavior. The concept has great potential to reduce the accident rate on roadways, especially for young inexperienced drivers who may be suffering from chronic health conditions that when uncontrolled, can result in dangerous driving actions. Our approach involves building models of "normal" and "abnormal" driving by an individual through machine learning from observation (MLfO, or simply LfO). Conceptually, discrepancies between actual driving actions taken by a driver in real time and the actions prescribed by a model of her/his normal driving, and/or similarities to a model of his/her abnormal driving, could indicate a dangerous medical condition. If appropriate, the system could alert the driver and/or the appropriate authorities (e.g., EMTs, police, or parents if a minor) of the potential for danger. More specifically, our research created models of human driving through the use of an LfO system developed previously in our laboratory called *Force-feedback Approach to Learning from Coaching and Observation with Natural and Experiential Training* (Falconet). Time-stamped traces of actions taken by 12 human test subjects in a driving simulator were collected and used to create the models of human driving behavior through Falconet. Then the overall actions prescribed by the models (called the *agents*) were compared to the original traces to ascertain whether similarities and/or differences between the human test subject behaviors and the agent behaviors could be indicative of the target conditions. In our use case presented here, the target condition was Attention Deficit/Hyperactivity Disorder (ADHD), a condition that afflicts many driving age teenagers and which can be detrimental to safe driving when not under control through medication. The work described in this paper is exploratory in nature, with the objective of showing scientific feasibility. The results of extensive testing indicate that the agents created with the Falconet system produced promising results, being able to correctly characterize traces in up to nearly 82% of the test cases presented. Nevertheless, as is typical in such exploratory works, we found that much further work remains to be done before this concept becomes ready for commercial application. In this paper we describe the approach taken, the agents created and the extensive quantitative experiments conducted, as well as any insights learned. Areas of further research are also identified and discussed.

* Corresponding author.
E-mail addresses: gonzalez@knights.ucf.edu (A.J. Gonzalez), Josiah.w.ucf@knights.ucf.edu (J.M. Wong), t.emilymarie@knights.ucf.edu (E.M. Thomas), aleckerrigan@knights.ucf.edu (A. Kerrigan), hastingsl@knights.ucf.edu (L. Hastings), Andres.posadas@knights.ucf.edu (A. Posadas), aswu@cs.ucf.edu (A.S. Wu), santi@drexel.edu (S. Ontañon), ylee65@gmu.edu (Y.-C. Lee), winston@chop.edu (F.K. Winston).

https://doi.org/10.1016/j.eswa.2022.117167
Received 15 December 2020; Received in revised form 22 July 2021; Accepted 31 March 2022
Available online 4 April 2022

1. Introduction

Driving a motorized vehicle on US roads - or anywhere in the world for that matter - can be dangerous to one's health. The US National Safety Council estimates that there were 38,800 traffic fatalities in the US in 2019[1]. While these numbers have not grabbed the attention of the press or of our political leadership in the midst of the COVID-19 pandemic, traffic fatalities remain a serious problem that takes a terrible annual toll in lives, and cries out for innovative ways to reduce these losses. The work we describe here seeks to bring Artificial Intelligence to bear to reduce this needless loss of life, especially in young drivers. The research is the result of a collaborative five-year program by four institutional partners: Drexel University, Children's Hospital of Philadelphia (CHOP), George Mason University (GMU), and the University of Central Florida (UCF). This article only describes the research conducted (mostly) at the University of Central Florida, but discusses the overall objectives of the larger project to which all partners have contributed in ways not discussed in this paper.

Our overall concept explores opportunities for identifying health issues in a person in real time while he/she drives a motor vehicle. Our objective is to build a general approach for detection of abnormal driving behavior. Such abnormal behavior may be influenced by active health conditions, such as driving under the influence of alcohol or drugs, or may be the result of an event (e.g., stroke or heart attack). Our use case particularly targets young drivers who may be afflicted with complex disorders whose identification strictly via driving actions may be complex. As our use case, we address the detection of uncontrolled Attention Deficit/Hyperactivity Disorder (ADHD), a prevalent chronic medical disorder that when not controlled, has the potential for known negative health and quality of life consequences, including motor vehicle accidents[2].

Many other similar works reported in the literature (see the related works in section 3 below) involve physically instrumenting the driver to monitor his/her biological signals such as blood pressure, heart rate, skin temperature, direction of eye gaze (Groom, van Loon, Daley, Chapman, & Hollis, 2015) (Karatekin, 2007), EEG (Groom et al., 2010 – although not on a driver) and others. Unlike these works, our approach does not involve instrumenting the driver and is therefore non-invasive. It only monitors variables that can be automatically accessed from the car, such the car's speed, acceleration/deceleration, angle of steering wheel, plus several environmental variables such as traffic signals and the presence of other vehicles around it. Much of the required data (e.g., speed, angle of steering wheel) can be obtained directly from the automobile. Many of the environmental variables, such as speed limit in road segment and proximity to traffic signs or signals are already easily available from modern road navigation systems. On the other hand, when in actual practice, the detection of other vehicles, pedestrians and road hazards will initially require complex instrumentation such as cameras and machine vision interpretation. However, in a (possible) future of inter-vehicle communication and highly instrumented road networks, especially in urban settings, such information may be available directly through these means, thereby making such additional instrumentation potentially unnecessary, or at least minimal.

Specifically, we sought to investigate how machine learning could be employed to detect abnormal driving behavior. We use the concept of *Machine Learning from Observation* (MLfO, or simply, LfO) of human performance as the basis for our approach. LfO is a type of machine learning that builds a model of a person's behavior (actions) strictly through unobtrusive observation of his/her behavior. We call these models *agents*, and they are capable of prescribing a control action to be taken in a just-in-time fashion as a reaction to the situation being faced by the driver (we call it the *context*), as perceived in real time through a suite of sensors. The overall behavior prescribed by the agent as it "drives" the route alongside the human driver is then compared to the human's actual behavior in real time to detect any meaningful similarities or discrepancies that could suggest a problematic condition. Our work uses an LfO system (described later) to generate two models of an ADHD-afflicted automobile driver's behavior: one under "normal" conditions (i.e., not under the influence of ADHD), and another one under "abnormal" conditions (i.e., under the influence of ADHD). This approach involves an extra step – that of building these models (agents) a priori. However, we believe that there may be advantages to using agents that can operate in real time in reaction to the traffic context, and that are generalized (i.e., can cover similar but not identical situations to what was observed during their training).

Randell, Charlton, and Starkey (2020) report that there is significant difference in driving behavior between the ADHD-afflicted drivers who are *medicated* and those who are *un-medicated.* We thus define "normal" behavior here as that of an ADHD-afflicted driver who is currently on a correct dose of appropriate medication, while an "abnormal" driver is an ADHD-afflicted driver who is currently not on any relevant medication and therefore under the active influence of ADHD.

The LfO system used to generate the two types of agents (under medicated and un-medicated conditions) was the *Force-feedback Approach to Learning from Coaching and Observation with Natural and Experiential Training* (Falconet) (Stein & Gonzalez, 2011; Stein, 2009). The agents are trained by Falconet to learn how to drive a car only by observing how humans do it. Upon completion of the training process, they are capable of determining a driving action in light of the driving situation, and executing it in a simulation. A subsequent extension of Falconet (Stein & Gonzalez, 2014) made it context-centric, such that it prescribed the actions based on the *context* identified through interpretation of the sensor readings; this context is to be the same as what the driver would perceive. This contextualization permitted agents created with this extension to succeed in applications where agents created with the original version of Falconet could not (see Stein & Gonzalez, 2014). Actually, with some modifications, these agents could conceivably be used to drive a physical car autonomously, but that is a story for another day.

The agents used in our research were created from unobtrusive observation of 12 human test subjects as they drove a simulated vehicle (called *own-car*) through a simulated road network. The use of a car simulator eliminated safety risks to the human test subjects, to our research staff, and to any other drivers and pedestrians who might have been present if we had gathered the data while driving an actual automobile on actual roads. This would have been particularly perilous when a test subject afflicted with ADHD was to be driving the car while un-medicated.

When realized to its fullest potential, we envision this approach to work as follows: an at-risk driver who suffers from a chronic medical condition is brought to a test center when in a normal state of health (i.e., medicated), and asked to drive a simulated automobile through several different traffic contexts. The same individual could be asked to return to the test center while un-medicated (on a different day) to drive the same simulated road network. Her/his performance data are collected as time-stamped *traces*, which are then presented to Falconet (after some non-trivial pre-processing) to create an agent that accurately reflects her/his normal (or abnormal) driving style. This agent is thereafter placed on-board his/her car on a computer with access to a suit of sensors, and is used to predict what the driver would normally do under the traffic conditions being perceived by the agent through the suite of sensors. Any discrepancies between the normal agent and the driver's actual actions - that are consistent over time and/or severe - could be flagged as indicative of an abnormal condition, and corrective action could be initiated by the system by contacting the appropriate authorities before an accident occurs. Additionally, a second agent that is reflective of how the same individual drives when under the influence of his/her chronic condition (an abnormal agent) would look for

[1] https://www.nsc.org.
[2] https://chadd.org/for-adults/adhd-and-driving/.

similarities between the abnormal agent and the driver's actual actions. Ideally, both could be used to obtain corroborating evidence. We should note here that our research described here did not include placing these agents in an actual automobile to detect the conditions in real time. We only assessed the after-the-fact correlation between an agent-generated trace and a human-generated trace, as the reader will see below. Placing the model on-board an automobile to detect abnormal conditions in real time is our logical next major step but it is regrettably left for future research.

2. Objectives of our research

The major objective of our research presented here was to show the scientific feasibility of our general approach to identifying negative health conditions strictly by monitoring a driver's driving actions. We use medicated and un-medicated drivers afflicted with ADHD as our use case to show such feasibility. Our hypothesis in this research is:

> Models of human driving behavior, built through machine learning from observation of human drivers afflicted with ADHD, can be compared to their actual driving behavior to identify when these drivers are operating a motor vehicle while suffering from un-controlled ADHD conditions.

Fundamentally, our approach looks for overall differences in driving actions over an extended (several minutes) period of time, rather than for specific actions that may be typical of drivers suffering from a specific condition (e.g., lane drifting, running stop signs or red lights, ignoring emergency vehicles). The advantage to this approach is that it could be generally applicable to many different conditions, without having to specify and codify specific behaviors for each condition. In such a way, we avoid the problem of a system that does not recognize a telltale behavior because it was not defined and codified a priori. In other words, our approach is unlike that of a knowledge-based system that could only identify specific behaviors known to an expert and properly codified into the system.

More specifically, we pursued the following two objectives:

1) Discern whether the state (e.g., medication condition) of a specific human driver can be identified by comparing her/his current driving actions to those of models of his/her driving behavior while medicated and/or while un-medicated. The question, in effect, is whether their normal and abnormal driving behaviors are sufficiently different and whether the actions that embody the differences are able to be captured and reflected sufficiently well in models built through machine learning from observation. If so, this would indicate scientific feasibility.
2) Assess whether the driving actions of several medicated humans are similar enough among themselves such that generic models of "normal" driving can be built for use on multiple humans. Similarly, the same was done for drivers in an "abnormal" state. If they are similar enough, conceivably one generic model could be used for all (or at least several) drivers in the same medication state. This would facilitate the exploitation of this technology in actual applications. Otherwise, individual models specific to each driver would have to be created.

We refer to the first as a *Horizontal Assessment*, where in our ADHD use case, the medicated and un-medicated agents pertaining to one individual test subject are compared to traces of the same test subject driving while medicated and while un-medicated. We refer to the second objective as *Vertical Assessment*, where a generic agent trained with multiple traces of different human test subjects driving under the same medication condition is compared to the traces of other test subjects in the same condition. The agents generated by Falconet were trained to output the speed of the car and the steering wheel angle. The speed models the foot pressure on the accelerator or brake pedals. The steering wheel angle situates the simulated car on the roadway and in the lane. Tests were designed to verify/disprove our hypotheses and objectives above. More details on these tests can be found in section 4.7 below.

We next describe the relevant literature to properly place our work within the state of the art.

3. Related work in driver health monitoring

In this section, we review the work of others who have investigated the impact of ADHD on adult and teenage driving behaviors as well as the concept of monitoring a driver with the objective of discovering some sort of physical or mental state on the part of the driver. We should note here that we do not claim to have made any advances in machine learning as a part of our research presented here. Therefore, we do not review any of the very extensive literature in machine learning.

There is ample evidence in the literature about the effects of ADHD on human task performance (Boland et al., 2020) dating back to the late 1990′s. Reimer, Mehler, D'Ambrosio, and Fried (2010) and Biederman et al. (2012) state that teenagers and young adults with ADHD "… have been shown to be at increased risk for impairment in driving behaviors" (Biederman et al., 2012). Barkley, Murphy, O'Connell, and Connor (2005) extend that statement to adults as well as teenage drivers. Curry et al. (2017), Curry, Yerys, Metzger, Carey, and Power (2019) take this further by reporting that ADHD-afflicted teens have a higher incidence of motor vehicle crashes in their first few months after initial licensure than do the general population of teenagers in their same first few months after licensure. Merkel et al. (2016) recorded videos of ADHD- and non-ADHD-afflicted drivers and had human judges observe recorded high G-force events and determined that drivers with ADHD exhibited more risky driving behaviors and increased consequences for "faulty driving". Fabiano et al. (2016), Faraoner et al. (2019) and Aduen, Cox, Fabiano, Garner, and Kofler (2019) provide recommendations on interventions to improve the driving habits of young drivers.

In a meta-study, Vaa (2014) refutes the mainstream opinion indicated above and asserts that the accident rate for ADHD-afflicted drivers was in fact not higher than for non-afflicted drivers. This report claims that the reason for the higher accident rates reported for ADHD-afflicted drivers in several other studies such as (Barkley, Guevremont, Anastopoulos, DuPaul, & Shelton, 1993) was that ADHD-afflicted drivers tend to drive more miles than those not afflicted. Nevertheless, we base our work here on the predominant opinion that un-medicated ADHD-afflicted drivers can present a greater driving risk than those who are medicated (or not afflicted with ADHD).

A significant amount of work has been done in monitoring the state of drivers, principally around detecting loss of driver consciousness resulting from fatigue/lack of sleep or intoxication. Das, Zhou, and Lee (2012) collected data from 108 drivers under normal conditions and under alcohol-induced impairment in a driving simulator. Based only on steering wheel movement, they differentiated driving states based on nonlinear invariant measures (such as sample entropy). Jin et al. (2013) detected sleepiness based on eye movement using machine learning (Support Vector Machines) in a study with 12 drivers. They report predictive accuracies between 74% and 96% when trained for specific drivers, and of 72% for a general model. Liang (2009) used data mining algorithms on a 100-car naturalistic dataset to detect driver distraction. Otmani, Pebayle, Roge, and Muzet (2005) reported on which driving performance measures were correlated with sleep deprivation. Kang (2013) applied several techniques to the detection of drowsiness and distraction, and to predictions of dangerous driving.

Research has progressed sufficiently for automatic detection of drowsiness such that a range of approaches exist based on in-vehicle, behavioral or physiological measures (Sahayadhas, Sundaraj, & Murugappan, 2012). Sahayadhas et al. review approaches based on detecting yawning, facial expression and eye-related measurements; head pose and gaze direction; approaches based on physiological measurements;

and those based on vehicle features such as steering wheel or lateral position deviation, e.g., (Torkkola, Massey, & Wood, 2008). They conclude that there is an urgent need for public datasets on real driving conditions to evaluate these techniques. Their research demonstrates the feasibility of using driving behavior to detect clearly dangerous driver states (distraction, intoxication and drowsiness) but the work has not extended to subtle (nonintrusive) monitoring of more complex behaviors resulting from medical conditions such as ADHD to predict when the conditions of the drivers are not well-controlled.

Other work has focused on predicting high-risk situations before they happen. Siordia, de Diego, Conde, Reyes, and Cabello (2010) present a machine learning approach to predict driving risk level, based on features extracted from a driving simulator (including visual features, such as "driver is not holding the steering wheel"), achieving classification rates of higher than 90% when the model was trained and tested in the same scenario, but only between 20% and 60% when they tried to generalize across scenarios, indicating that the sets of rules (obtained from experts) used for preparing the data are highly domain dependent. More robust models are needed for this important predictive capability.

There has been work on learning predictive models of driving behavior. Suzuki et al. (2005) propose a model to learn driving behavior (with focus on collision avoidance) based on stochastically switched linear models. Specifically, they learn linear models, and then learn a switching policy based on Hidden Markov Models (HMMs). They observe that often a human driver "switches the simple control laws instead of adopting the complex nonlinear control law". Oliver and Pentland (2000) presented an approach based on HMMs and Coupled HMMs (CHMMs) trained from data from a Smart Car from 70 different drivers, that was capable of predicting behavior up to 1 s before the maneuver takes place. Kishimoto, Abe, Miyatake, and Oguri (2008) present a Dynamic Bayesian Network model capable of differentiating hasty and un-hasty behaviors. These models were trained separately for different maneuvers: passing, starting and stopping, turning left, etc. Salvucci (2004) presents an ACT-R-based model for predicting lane changes.

Finally, other members of our project group have addressed the same problem from different Machine Learning perspectives. Our partners at Drexel University use clustering methods and classification algorithms (k-NN) to identify segments in a time series that show similar traits (Grethlein & Ontañón, 2020; Grethlein et al., 2020). Likewise, our other partners at George Mason University employ a direct comparison approach that selects specific regions of the trace and creates a contrast dynamic feature dependency (CDFD) graph or pattern of the event (Li, Zhao, Lee, & Lin, 2020; Li, Zhao, Lee, Sassanin, & Lin, 2020).

4. Our approach

In this section, we briefly describe the LfO system used to create the agents, the data used for creating these agents, the simulator used to capture these data from the human test subjects, and the tests conducted. We begin by describing the underlying technology used to create (train) and execute our agents.

4.1. Machine learning from observation

The process of machine learning from observation creates agents that learn to act in a manner similar to a human actor who is observed as she/he performs a control task (e.g., controlling some type of physical device). The observations are recorded in a time-stamped trace, taken either from a simulator (as in our case) or from an actual device in the physical world. Two important (self-imposed) restrictions on our work are that: 1) no interaction of any type is permitted with the performing actor, other than the quiet observation of her/his behavior. That means that no questions can be asked of the actor before, during or after the performance. Moreover, no further performances can be arranged to clarify any unresolved ambiguities. We refer to this as *unobtrusive observation*. Secondly, 2) the actor's performance is not for the purposes of training anyone or anything. Rather, it is to be an uninhibited natural performance of carrying out the task to be observed. These two restrictions allow the application of our techniques to observing actors who may not be aware that they are being observed (e.g., observing the tactics of enemy forces in a battle or of an opponent in athletic competition). These are the main distinguishing features between our version of LfO and *Learning from Demonstration*, an otherwise similar area of research.

There is an extensive body of research literature on Machine Learning from Observation, Learning from Demonstration and other variations of this theme that use labels such as *Behavioral Cloning* and *Learning from Instruction.* All of these share the common basic objective of learning through direct interactions with humans. Following up on our statement in section 3 above, we likewise do not claim here to have made an advance in LfO. Rather, we present a novel application of LfO systems and methods already described in the literature (i.e., Falconet). Therefore, we have refrained from providing a review of the extensive LfO body of literature in order to limit the length of this article. We refer interested readers to Argall, Chernova, Veloso, and Browning (2009) and Torabi, Warnell, and Stone (2019) for detailed reviews of the literature in these topics.

Our research group has done extensive work on learning from unobtrusive observation of human behavior since the early 1990s, mostly while driving an automobile or performing similar control tasks that have included driving a battle tank, tactically navigating a submarine, flying a drone, loading boxes on ships with a movable crane, and even herding sheep. The observations all took place in simulators to avoid the impracticality of real world exercises (e.g., finding a battle tank, a submarine or a herd of sheep), not to mention avoiding the aforementioned inherent risks to all personnel involved when driving real automobiles on public roads. The Falconet system used in our work had been previously applied to automobile driving behavior with good results (see Stein & Gonzalez, 2011), so it was our natural initial choice. Another LfO system that had been previously (and successfully) used to learn automobile driving behavior was the *Genetic Context Learning system* (GenCL) (see Fernlund, 2006) which used Genetic Programming (Koza, 1992) as its basis for learning. However, in an early phase of this investigation, our results with GenCL were not deemed to be satisfactory, so we desisted on its further use. Other potentially applicable systems such as Sidani's IASKNOT (Sidani & Gonzalez, 2000), Stensrud's FAMTILE (Stensrud & Gonzalez, 2008), Trinh's COPAC (Trinh & Gonzalez, 2013), Johnson's COLTS (Johnson & Gonzalez, 2014) and an un-named system by Aihe (Aihe & Gonzalez, 2015) had not been specifically or extensively applied to learning driving behavior and thus were not explored. We hope to study the applicability and usefulness of some of these other systems in our future research.

Falconet employs a novel algorithm called *Pigeon-Alternate* (simply called *Pigeon* here) (Stein, Gonzalez, & Barham, 2015) that combines Neuro-evolution (Stanley & Miikkulainen, 2002) and Particle Swarm Optimization (Kennedy & Eberhart, 1995) in an alternating manner. We discuss Falconet and Pigeon in greater detail in Section 5 below. Nevertheless, for full descriptions of Falconet and of the Pigeon algorithm we direct the interested reader to (Stein and Gonzalez, 2011, 2014; Stein et al., 2015) and particularly to the doctoral dissertation that served as source documents for these publications, (Stein, 2009).

4.2. Driving simulator

Our experiments comprised observing the driving behavior of 12 different human test subjects on a driving simulator. The data collection was done by our partners, the Children's Hospital of Philadelphia and George Mason University. Each of the test subjects was medically certified to be afflicted with ADHD. Each test subject was asked to drive a simulated automobile through four simulated and pre-determined routes (called *Drive 1* through *Drive 4*) while under two separate

medication states (or *conditions*). The test subjects drove each of these routes twice – once while under a *medicated* state and once while *un-medicated.* The drives under each medication state were done on different days. The routes presented the driver with traffic lights that changed color (red to green only), stop signs, and construction zones, speed limits, and included other traffic. The test subjects were allowed one unrecorded drive of arbitrary duration (Drive 0) prior to being recorded in order to allow him/her to become familiarized with the simulator.

The data used in our particular study were collected at George Mason University in a half-cab Realtime Technologies, Inc. motion-based high-fidelity driving simulator. The driving scenarios were programmed using JavaScript while the driving environment was developed in SimVista and run using SimCreator. Participants completed the practice (unrecorded) drive and the four different experimental (recorded) drives, each lasting between 7 and 15 min. The drives contained ambient traffic and consisted of two- and four-lane roads in rural and urban environments. Three cameras recorded the participants' foot movement, face, upper body, and over the shoulder view; however, these camera-based data were not used to build our models. See Fig. 1 for some visual scenes in the driving simulator, and Figs. 2 and 3 for simplified maps of the four drives with the traffic signals indicated therein. Table 1 describes the traffic devices present in each of the drives. See Lee et al. (2018) for more information about the study design, including details about the participant profiles, recruitment procedure, etc.

4.3. Agents created

Twenty-four (24) agents were created from twelve (12) human test subjects who were known to suffer from ADHD – 12 agents that model each test subject's individual behavior while driving medicated, and 12 other agents that model their behavior while driving un-medicated. The test subjects were specifically recruited for this study and had to pass a set of clinical screening criteria to assure ADHD affliction. They were anonymously labeled as Subject #602, Subject #607, Subject #608, Subject #609, Subject #612, Subject #613, Subject #615, Subject #617, Subject #619, Subject #620, Subject #622 and Subject #627. The same traces of these twelve human test subjects were used for the Horizontal and Vertical tests.

We should note here that although the test subjects drove predetermined routes (Drives 1 through 4) in their experiments, the Falconet-based agent building process in no way used this predetermination to assist in its creation of the agents. Thus, it can be safely asserted that the process is a general one and would work the same way with any other route driven as long as it had the same elements (e.g., traffic lanes, traffic lights, stop signs, etc.).

4.4. Context and its role in our approach

It is widely accepted in cognitive psychology that humans rely heavily on context for memory recall, speech, and problem solving (Hollister, Gonzalez, & Hollister, 2019). However, context had originally been an underappreciated component of artificial intelligence in the early AI literature (Zibetti, Hamilton, & Tijus, 1999). That has changed over the last 20 years or so since the emergence of context-sensitive computing (Dey, 2001). We are strong believers in the power of contextual reasoning for building intelligent agents, so the agents created in this project were designed to operate in a *context-centric* manner. That is, an agent is trained on how to act in a particular situation (a *context*) where certain specific assumptions are valid. An agent thus not only has to learn what to do when in a particular context, but must also be able to infer in what context it finds itself at all times so it can autonomously "place itself" in the correct computational context when the situation around it changes, so it can act appropriately. To infer the context in which it is, the agent needs to be able to recognize the environmental and internal conditions, as well as its own goals. To implement such context-centric behavior, the resulting agents operate under the *Context-Based Reasoning* (CxBR) paradigm (Gonzalez, Stensrud, & Barrett, 2008).

Briefly, CxBR dictates that an agent can be in one of several mutually-exclusive *Major Contexts*, which are situations where only a (usually small) subset of the agent's total knowledge is relevant. In CxBR, exactly one Major context, the so-called *active context*, controls the agent at any one time; so, it is imperative that the active context correctly reflect the situation in which the agent finds itself at all times. Each Major context has one or more *transition rules*, which determine to which other Major context the control is to switch when the active Major context is no longer deemed suitable to address the situation at hand. Transition rules, of course, are also context sensitive so that the Major context to which to switch depends on the active context. It also contains one or more *action rules* (or *action functions*), which determine what the agent should do when it is under the control of this context. However, when a relatively brief and specialized situation presents itself, a context can temporarily pass control of the agent to a *sub-context* designed to handle that specialized situation. An example of such specialized situations could be passing another car on a two-lane road. In this example, while driving on a rural two-lane road and being controlled by a Major Context appropriately called **Two-lane-road-driving**, the agent encounters a slow moving truck ahead and decides it wants to overtake it. The Major (active) Context passes control of the agent to a sub-context that might be appropriately called **Passing-in-two-lane-road**, which manages the agent's actions while it is passing the truck, only to return control to the original active Major Context that invoked it after the passing procedure is completed. However, if the agent's situation were to change (at any time) such that the active context is no longer suitable to address it, then the active context's transition rule(s) will return false and a new Major Context, whose (at least one) transition rule(s) return true, would become the new active context. If no Major Context is able to take control (should only happen very rarely), then a *default context* becomes the active context.

In our work, each agent was composed of five independent context agents – Rural Construction Zone (RCZ), Rural Stop Sign (RSS), Urban Construction Zone (UCZ), Urban Stop Sign (USS) and Urban Traffic Light (UTL) (see Fig. 4). These five context agents reflect the behaviors observed in the test subjects when they were in these different traffic situations. We should note that a CxBR agent normally possesses the knowledge to determine in which context it finds itself; however, such was not put into effect in our agents, as our intent was to assess whether the contextualized behaviors were accurate representations of the test subjects observed. Furthermore, while Falconet possesses the means to automatically learn the transition rules that serve to trigger context changes, these features were not activated in our work. So, when

Fig. 1. Scenes (from left to right): Drive 1 - approaching stop sign; Drive 2 - cruising down a straightaway; Drive 3 - navigating construction zone; Drive 4 – approaching a light.

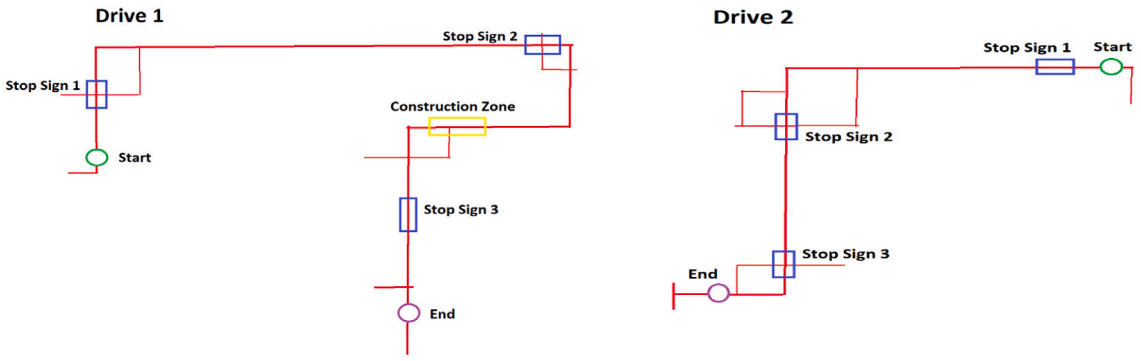

Fig. 2. Simplified Depiction of Drives #1 and #2.

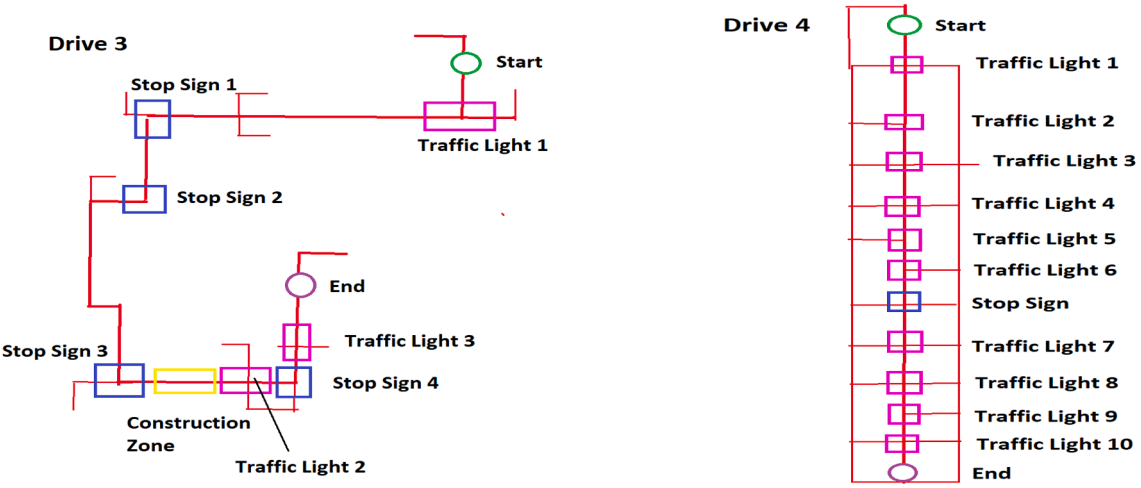

Fig. 3. Simplified Depiction of Drives #3 and #4.

Table 1
Driving scenario characteristics.

Drive	Traffic Light State		Inclines		Curving Turns		Other		
	Green	Red→ green	Uphill	Downhill	Right	Left	Stop Signs	Hazards	Length (m)
Drive 1	-	-	5	5	3	1	3	5	12,000
Drive 2	-	-	1	4	1	1	3	1	6,700
Drive 3	3	-	-	-	2	1	4	1	6,300
Drive 4	-	10	-	-	-	-	1	1	3,700

evaluating the performance of an agent, its different contexts were compared only to the corresponding context segments that were manually identified in the drives (and used in the training). We therefore leave for future research the automated transitioning between contexts, something that will ultimately be necessary for practical implementation.

4.5. Data used for training the agents

Our observation of the test subjects driving the simulated car (own-car) consisted of recording forty (40) variables (e.g., speed, acceleration, deceleration, location, nearby traffic, etc.) as a function of time and at a frequency of 10 Hz, throughout each test subject's four drives. The resulting time-stamped record of these variables and their values over an entire continuous drive (the traces) of driving behavior, constitute the record of our observation of human driving behaviors.

Seven of the 40 recorded variables reflect the relative location of own-car: these are the distances to: (1) the nearest stop sign, (2) a construction zone, (3) a traffic light, (4) a left turn to be made at the nearest intersection, (5) a right turn to be made at the nearest intersection, and (6)-(7) left and right curving turns respectively. These variables serve to gauge the distances to static road objects or locations to which a driver may react as he/she approaches them.

Two other variables describe other aspects of the simulation: (8) simulation time, and (9) road slope.

Twelve (12) more variables were recorded that measure the following for own-car: pedal pressure (10) for throttle and (11) for brake; (12) speed; (13) acceleration; (14-15-16) XYZ coordinates of own-car; (17) cumulative distance traveled in 2D; and (18) cumulative distance traveled in 3D; (19) steering wheel angle; (20) own-car heading; and (21) heading error. These variables measure how the driver controls own-car, in reaction to the immediate environment and its

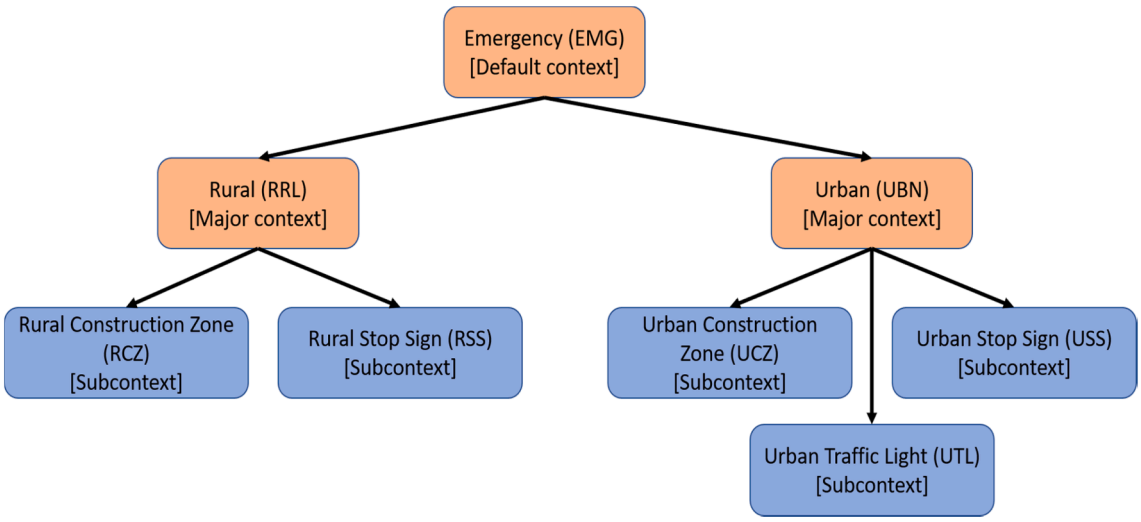

Fig. 4. Context hierarchy for controlling driving agents.

effect on the car.

Finally, there are sixteen (16) variables that track other vehicles (OVs). OVs are a highly dynamic element of driving and a crucial consideration in road safety. Four OVs can be tracked at a time, one for each "ray" region around own-car (see Fig. 5). The area within a 300 m radius of own-car is divided into frontal and rear rays (Rays 1 and 3, respectively, covering 60 degrees each) and left and right rays (Rays 2 and 4, respectively, covering 120 degrees each). The closest OV in each ray (if any) is tracked. For each of these four OVs, we track their: X-Y coordinates; distance to own-car; and approach speed relative to own-car. These make the total 37.

Lastly, the color of a traffic light, the speed limit in force in the road segment being traveled by own-car and the lane offset were also acquired from the simulation, but they were in the form of one hot encodings. The lane offset has six possible values, only one of which is set to true (1) at any time. These assume a four-lane road with no median, and are: "off-road left", "left outer lane", "left inner lane", "right inner lane", right outer lane" and "off-road right". Likewise, traffic light color can be one of three values: red, green and none (a yellow light is never observed in our simulation). Lastly, only four specific speed limits are possible: 3, 15, 35 and 45 MPH. Counting these as three discrete variables (rather than 13 one hot encodings), makes the final variable count 40.

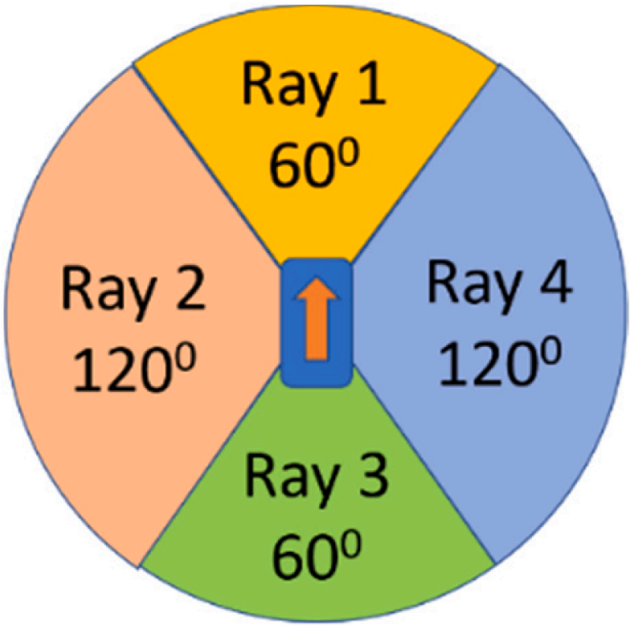

Fig. 5. The ray regions around a car to track other vehicles (not drawn to scale).

4.6. Training data preparation and presentation

The traces of human driving behavior were collected by our partners at George Mason University as discussed in section 4.2. These traces were then subjected to a data preparation process to enhance the traces with additional information, correct certain variables, and modify the trace format to be usable by Falconet.

First, a data pre-processing module written by our partners from Drexel University was used to compute and add some variables, such as the distances and relative speeds to the various OVs.

Secondly, another pre-processing module was written by the UCF team that made the following modifications to the data on the traces:

1. The sampling rate in the traces was downgraded from 60 Hz to 10 Hz. This was done because, in our opinion, the 60 Hz sampling rate was unnecessarily high, as events do not happen that quickly in automotive traffic actions. Moreover, it allowed our LfO algorithms to learn from longer training segments without being overwhelmed by useless data.
2. The distance to the closest urban construction zone was added.
3. The distance to the closest traffic light was modified to only consider traffic lights that govern own-car's lane of traffic.
4. The values of each variable were normalized to values in the range [-1.0, +1.0] to make them usable by neural networks (i.e., Falconet).
5. Converted the traces to regular text files.

The raw (unedited) traces recorded by the simulator provided the car's speed in meters/sec. Expanding upon item 4 above, these speed values were subsequently normalized (as were all data) before being presented to Falconet The same min/max values were applied to every trace of every test subject and were selected to encompass the min/max values observed over all traces. Not all normalized values were in range $[-1.0, +1.0]$, however. If a variable's min value was not <0, then the normalized values for that variable was in range [0, 1] instead. Thus, for a range defined as [m, n], given a value V and min/max values A and B, the normalized value D was computed as:

D = m + [(V-A)/(B-A)]*(n-m).

For speed, the min/max values used were −10 and 30 m/s, which is equivalent to −22 and 67 miles per hour (it is possible for the human to drive backwards).

For steering angle, the min/max values used were −600 and 600 degrees, where 0 degrees is dead straight ahead. As done for speed, the normalized steering angle values presented to Falconet were in range [−1.0, +1.0]. A value of 360 degrees in absolute value corresponds to turning the wheel one full revolution (in either direction). Thus, −1.0 is 1.6 revolutions counter-clockwise, 0 is dead ahead, and 1.0 is 1.6 revolutions clockwise.

Third, once the raw traces had been captured, enhanced and normalized, the next step was to de-compose these traces into contexts - the *contextualization* process - to present them to Falconet for training the agents. Contextualization involved manually partitioning each test subject's enhanced and normalized trace into segments belonging to particular context instances (there may be several instances of the same context in each drive). The approach taken to partitioning the context instance segments involved some hard rules as well as some human judgment – "eyeballing" if we may. The locations of stop signs, construction zones and traffic lights were well-known from the drives used. So, these were the hard rules. However, not so obvious was exactly how far before a stop sign would a driver enter the stop-sign context, nor how soon after driving past the stop sign would one leave it. The same was true for the traffic lights and the construction zones. In the case of stop sign and traffic lights, the number of cars backed up behind it would also influence when the context would begin. The latter decisions were made judgmentally by the researcher performing the data preparation, and his/her judgment was facilitated by a tool called the *Traffic Simulator* that graphically displays the location of the moving cars and the traffic elements. See Wong, Hastings, Negy, Gonzalez, Ontañón, and Lee (2018) for details. The Traffic Simulator was an in-house tool that used OpenGL graphics to represent cars and road entities with simple colored geometric shapes and lines. The tool was used to replay a human's behavior in the road environment, allowing a researcher to easily determine when the human's behavior changed in reaction to an external stimulus, thereby possibly denoting the start or end of a context of interest. Training segments tended to be anywhere from 2 to 20 s in duration.

Fig. 6 shows a screenshot of what the user would see in the Traffic Simulator. It illustrates an example of part of Drive #4 as the driver approaches the first group of cars (OV's). There is a traffic light (currently red), and indicated by the circled red square on the side of the road where the state of the traffic light (i.e., its color) would be seen. Also shown are two event triggers that have already been triggered; this is indicated by the faded red squares, which means the driver has passed over them and in these situations, they triggered the other cars to appear ahead of the driver.

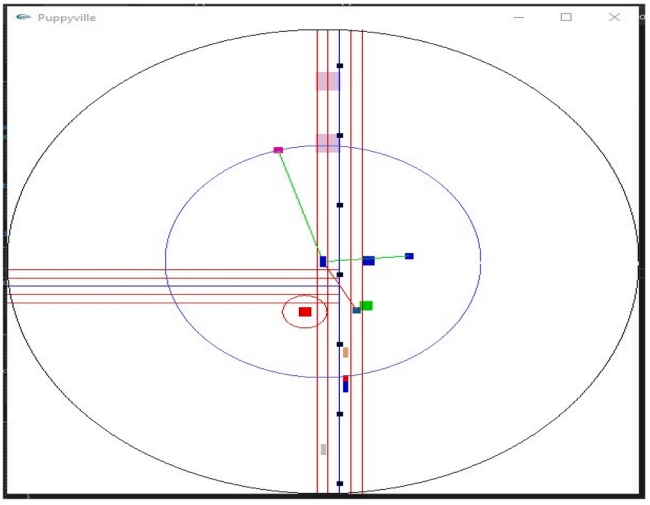

Fig. 6. Output of Traffic Simulator showing Drive #4 (mostly urban).

Fourth, once the start/end time steps of several training segments denoting a given context (e.g., rural stop sign) were determined by the researcher, a script was applied to format these training segments into XML files that were used by Falconet. In order to ensure that each training segment exerted the same influence on agent training, regardless of the training segment's temporal duration, each training segment has approximately the same number of comparison points - the points at which an agent's performance was compared to that of its human test subject in order to determine its fitness. The optimal number of comparison points, as well as the number of time steps between comparison points for each segment, was computed programmatically via a script and stored in XML format.

To train the individual agents used in the Horizontal tests, Falconet was presented with four subsets of the main trace that reflect instances of contextualized segments (for each context) taken from the four Drives driven by the test subject. Falconet averages the values of the four traces at each time step in the simulation and uses this to calculate loss for training. The Generic agents used in the Vertical tests were similarly trained, except using 12 such sub-traces for each context (four from the main traces of each of the three test subjects whose behaviors are reflected in the Generic agents).

4.7. Description of tests conducted

In this section, we shed additional light on the tests that were performed as part of our assessment.

Our tests involved subjecting the agents created by Falconet from observed human-generated traces to the same drives driven by the test subjects. These agents are expected to perceive the simulated environment and produce an output consisting of the car's overall speed, and the angle of the steering wheel at a given time step. These are the two main variables that can be directly controlled by a driver. These data generated by the agents are recorded in another trace – the *agent-generated trace*. The output variables of interest in the human-generated trace were continually compared to the same output variables in the corresponding agent-generated trace throughout the entire drives. We elected to use the agent's (and the human's) speed rather than their throttle/brake pedal pressure output because speed is a more reliable variable for post-hoc evaluation of the created agents trained in vehicular simulations. The reason for this, as noted by Fernlund (2006), is that the automobile (or a simulation thereof) acts as a low-pass filter when converting throttle/brake pedal pressure into speed. Even though throttle/brake pedal pressure may experience much variance within short spans of time, only sustained pedal pressure will result in meaningful and externally observable fluctuations in speed.

In addition to speed, the angle of the steering wheel plays a crucial role in vehicular navigation. Unlike in straight driving, where one must simply keep the steering wheel straight and maintain speed, a turning maneuver requires coordination between the steering wheel and car speed. If a driver makes an adequate turn of the wheel but is going too fast, the car may flip over, or if the driver turns the wheel too slowly, the car may collide with other traffic. Therefore, comparing speed and steering angle combined was of primary interest in our research.

The objective of comparing a human-generated trace to its corresponding agent-generated trace was to determine whether the driving behaviors reflected therein were similar or divergent. Similar traces indicated that the human driver's actions were generally correctly predicted by the agent, while divergent traces indicated the inability of the agent to predict the actions of the driver as reflected in his/her trace. As mentioned earlier, this drives to the crux of our research – determining whether a driver is driving normally or not.

Which agent-generated and human-generated traces were used in these comparisons, of course, depended on what the specific comparison

was intended to assess. This is where the difference between Horizontal and Vertical tests comes in. The former sought to determine differences between driving behaviors of the same driver in various health states. Specifically, if a driver's agent modeling his/her normal behavior were to indicate divergence with his/her actual driving actions, it would indicate that he/she is being influenced by some factor. Therefore, Horizontal tests can be used to detect abnormal driving. Success in the Horizontal tests is defined as when the behavior of an agent compared to its corresponding test subject trace of the same state (i.e., *co-comparison*) is significantly more similar than when compared to the same test subject but in the opposite state (e.g., *cross-comparison*). In practice, this would imply that a driver's medication state could be inferred by seeing how well her/his driving behavior matched that of its agent(s).

Vertical tests, on the other hand, seek to determine uniformity among drivers and agents who are in the same state of health. This is important because if reasonable uniformity can be shown, then there would only be a need for a generic agent that could represent a large number of drivers in the same state of health, rather than one agent for each individual driver. Success in Vertical testing would occur when there is strong similarity among the several human traces compared to the generic agent that reflects the same state of medication.

The terms "horizontal" and "vertical" came from the idea that if the agents and the human traces were to be listed in two side-by-side stacks, horizontal tests would only compare each agent to the trace directly across from it (pertaining to the same test subject), while vertical would compare the agents to several traces on the other stack, thus appearing more vertical. Fig. 7 displays the difference between Horizontal and Vertical tests. The solid arrows in the Horizontal tests on the left side of the figure reflect the *co-comparisons* while the dashed arrows represent the *cross-comparisons*.

Fig. 8 displays a graphical depiction of the hierarchy of the tests conducted on the agents.

The test used for all our assessments was the *Learning Capabilities* (LC) test, which compares the behavior of a trained agent against the trace used to train it and against the trace of the opposite medical state of the same test subject. So, it measures how well the agent learned from the human test subjects, and how different that is from the behavior of the subject in the opposite medication state. The results currently reflect the agents being trained on the five contexts mentioned above: USS, UCZ, UTL, RSS, and RCZ.

The discussions above beg the questions: How do we determine similarity or divergence? What metrics do we use to make an informed judgment on similarity or divergence? The next section discusses these questions.

4.8. Metrics used in the assessments

We considered several possible metrics for our assessment. We elected to use the Pearson Correlation Coefficient as a good way to measure similarity in terms of correlation of variables over entire road segments that were partitioned according to the contexts involved in each segment. While other metrics could have also been suitable, we thought that having one single number – the Pearson coefficient – that could provide indication of overall similarity or divergence would best suit our objectives.

Formally speaking, the Pearson Correlation Coefficient (PCC) is a measure of the linear correlation between two variables over several points, in our case, time. We used this method to measure the linear correlation between two variables in two traces – in our case the speed and steering angle of the cars in the human-generated trace were compared to the same variables in the agent-generated trace. The Pearson correlation coefficient assesses how well variables in these two traces correlate with each other over the entire duration of the drives. The PCC is in the range $[-1, +1]$, where values close to $+1$ indicate high positive correlation, values close to 0 indicate no correlation, and values close to -1 indicate high negative correlation (doing the opposite thing). The formula for the Pearson Correlation factor is:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{y})^2}}$$

where X and Y represent the time steps of the human-generated trace and of the agent-generated trace for the same drive, respectively.

Because of our two-output Speed + Angle tests, the Pearson Coefficient computation had to be handled somewhat differently; this is instead of just one set of two variables as in the classical Pearson correlation equation above. Thus, the computation had to incorporate the differences between the time-varying values of the speed (human-generated vs. agent-generated) with their respective means, and the differences between time-varying values of the steering wheel angle (human-generated vs. agent-generated) with their respective means. We took the sum of these differences to represent the $X_i - X$ and $Y_i - Y$ terms in the equation above. Keep in mind that the values outputted by Falconet are already scaled from -1.0 to 1.0, thus avoiding any swamping by one variable over the other.

Once we have generated a value indicative of the degree of similarity between two behaviors, we now need to be able to identify similarity vs. divergence in the overall traces. The classification accuracy is an indication of this most important outcome of the tests - the overall correct identification percentage rate for each test subject/agent combination in each context. In other words, how often did the comparison between an

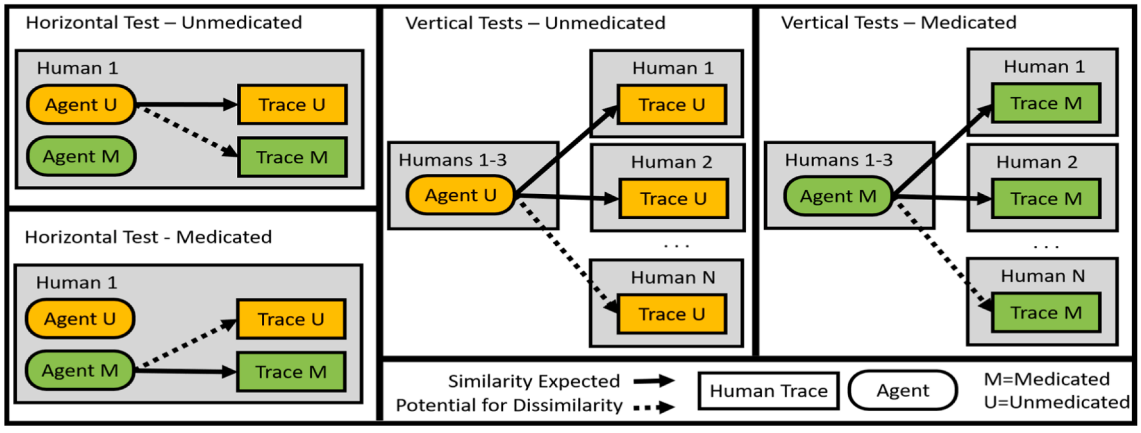

Fig. 7. Horizontal and Vertical Tests.

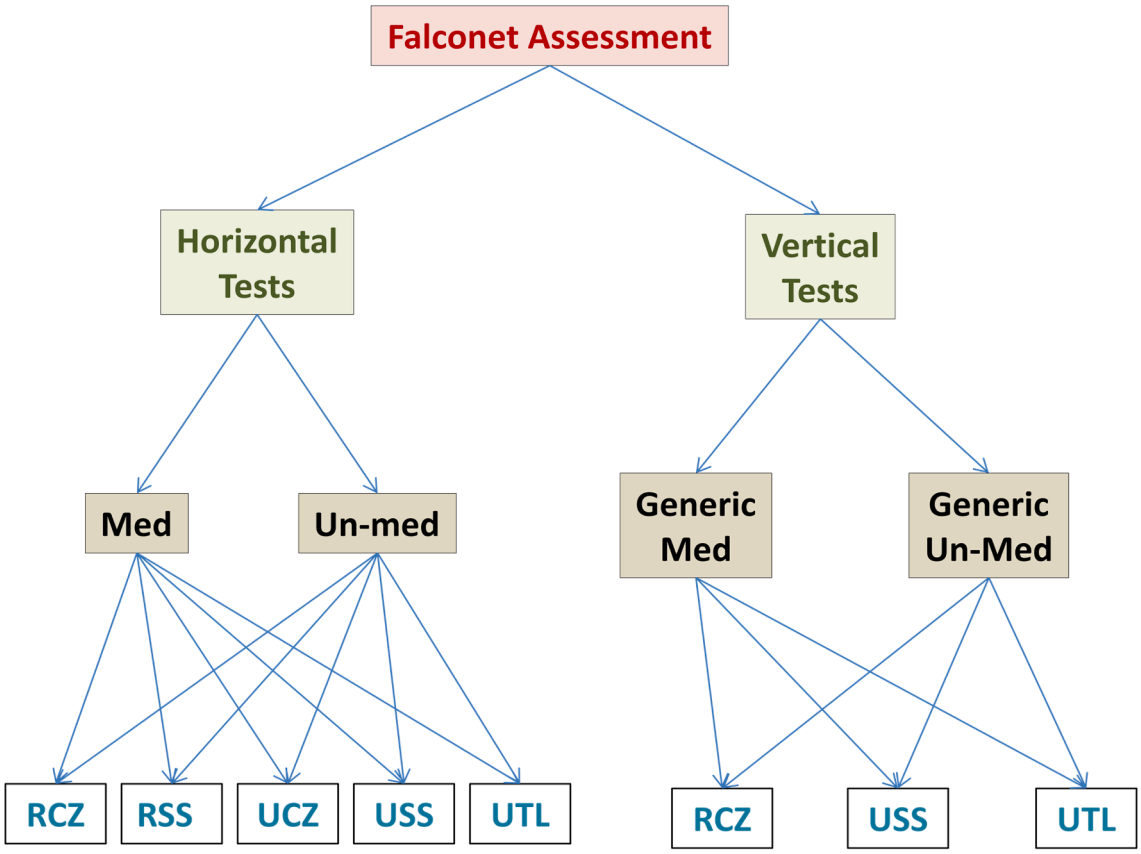

Fig. 8. Graphical Depiction of the Assessments Executed on the various agents.

agent and its corresponding two test subject traces result in correctly identifying the state of the test subject as reflected in her/his trace? This measure represents the bottom line of our work. The definition for a successful test result (a "*win*") is:

a. The agent correlated better with the human trace of the same medication condition than with the trace reflecting the opposite condition. We refer to these as "*positive comparisons*"
b. Only those positive comparisons deemed statistically significantly different were counted as wins. The p-value was computed for a 95% confidence factor to determine the statistical significance of any computed differences.
c. Furthermore, the absolute correlation coefficient for an agent must be greater than or equal to + 0.10 when compared to its human-generated trace of the similar medical condition. Otherwise, it is not counted as a win in our scoring, even if the correlation is significantly higher than that of the trace of the opposite medical state. This was done to avoid rewarding particularly poor correlations.

Conversely, unsuccessful test results ("*losses*") were those that:

d. Did not indicate better correlation with the trace of similar medication condition (i.e., "*divergent comparisons*").
e. Reflected a positive comparison but with a statistically-insignificant difference.
f. Were otherwise positive comparisons but the absolute correlation of the higher one did not reach the + 0.10 threshold correlation as described above.

The composite average of each agent over the twelve test subject traces was also computed on a context-by-context basis. While not as granular as the correctness metric above, it can serve to validate the overall correctness.

For vertical tests, we created a *Generic Agent* trained with a trace composed of a combination of the traces of three human test subjects in similar medication conditions. Therefore, there were two Generic Agents – one medicated and one un-medicated. The Generic Agent was compared to the three traces that composed its training data as well as to the other nine that did not. Success in these tests would be to show high correlation with all test subject traces. This would indicate that humans drive similarly while in the same medicated state. The implication of this is that generic agents can be built for this application rather than having to create two individual agents for each driver (medicated and un-medicated).

5. Description of Falconet

In this section, we provide a closer look at the Falconet system whose use was the central focus of our approach. However, to not repeat what has already been extensively published, we refer interested readers to (Stein & Gonzalez, 2011) (Stein & Gonzalez, 2014) (Stein et al., 2015) and especially the source document (Stein, 2009) for an in-depth description of Falconet and its various applications.

The heart of the Falconet system is the Pigeon-Alternate algorithm developed by Stein as part of his doctoral dissertation (Stein, 2009). It combines Neuro-evolution (Stanley & Miikkulainen, 2002) and Particle Swarm Optimization (PSO) (Kennedy & Eberhart, 1995) to build agents. Both processes are rather complex in their own right. Neuro-evolution involves its own combination of Genetic Algorithms (GA) and Artificial Neural Networks (ANN). While the literature contains many reports about the use of GAs to set the weights in ANNs, Neuro-evolution uses GAs to set the weights **and** the structure of the neural networks. An initially random *population* of relatively simple neural network

individuals is created and each individual network is trained with a subset of the examples available. It is then evaluated for its *fitness* as to how well it solves the problem at hand (or more specifically, how well it produces the desired outputs for the examples).

The fitness of each individual neural network solution (or just "individual") in the population is computed by executing its neural network and determining the success of the action taken by that individual in a brief simulation. Then, parts of those individual solutions judged to be more "fit" are combined or *mated* to form new solutions that are hopefully better. This process is called *crossover*. Finally, the individuals in the new population of solutions are randomly modified or *mutated* to introduce potentially beneficial solutions into the population. The mutation and crossover processes in Neuro-evolution systems are rather complex and their description is beyond the scope of this paper. The interested reader is referred to Stanley and Miikkulainen (2002).

The crossover and mutation processes are performed on some individual neural networks in the population, and the fitness is computed again. Only the better performing individuals are carried over into the next generation, thus implementing the survival of the fittest concept upon which genetic algorithms are generally based. As the ANN individuals are modified by the mutation and/or crossover processes, their structure becomes more complex, with additional levels, nodes and connections. This process is called the *complexification* of the individual neural networks, and is a central feature of Neuro-evolution. As more complex ANNs evolve, the complexification process adds new examples to the set used to train the ANNs, thus provoking the complexification.

The Particle Swarm Optimization technique is a non-linear stochastic optimization process. It is a variation of GAs that treats the particles "… like social groups with attractors, and the combination of individual agents can produce complex emergent behavior." (Stein, 2009, p. 75). At its most basic level, each of the particles is placed in a location in n-dimensional space. Initially at random, this location reflects the output values that serve to solve the problem, and are used to compute the fitness of the individual. Each particle has a small amount of memory in which it stores the fitness and state of the best that it (itself) has ever been in terms of these locations in the problem space. The particles seek to move toward the location that represents the best they have ever been but they also want to move toward the location of the best particle in the group. So, the particle computes a vector comprising these two locations and moves in this direction with a certain speed. The fitness of each individual is recalculated and the process begins again after it has moved along the computed vector in the previous generation.

The Pigeon algorithm works by executing Neuro-evolution for several generations, then interrupts the Neuro-evolution and applies PSO for several more generations to optimize the weights in each individual ANN. Pigeon resumes the Neuro-evolution process to begin the next cycle of neural network complexification. Pigeon continues to alternate (and thus its full name, Pigeon-Alternate) between several generations of Neuro-evolution and PSO until the objective is reached, there is no further improvement in the fitness of the best ANN individual, or it reaches a maximum number of generations. Upon completion of the execution, the individual with the best fitness value becomes the trained agent.

When Falconet begins training agents, it loads the test subject's trace, information about road scenarios, and the training set XML files. For each training segment for the context on which the agent is being trained, the appropriate portion of the trace is loaded as determined by the training segment definition in the XML file. Each individual in the population during a given neuro-evolution generation is simulated over all training segments via a *Micro Simulator*, which is a lower-grade approximation of the Traffic Simulator. The Micro Simulator was created to only approximate certain aspects of the Traffic Simulator that are inconsequential over a short time period in order to allow faster evaluation of individuals during training. The individual's state in the simulator is synchronized with that of the human at the start of the training segment, and then the individual is allowed to progress for a few time steps in the Micro Simulator (0.2 to 2.0 s, depending on the training segment). When the execution of an individual reaches a comparison point, its deviation from the human at the same point in the training segment is computed and the deviation is added to the individual's fitness. The individual is resynchronized with the human-generated trace and the process repeats until the end of the training segment. An individual's fitness is equal to its average deviation from its human "trainer" over all training segments; a lower fitness value means better learning performance by the individual.

6. Test results and findings

We are particularly interested in executing and analyzing the Horizontal tests, as these are the ones that will indicate ("prove" may be too strong a word in this case) whether our concept is scientifically feasible or not. In section 6.1 we describe the results of the Horizontal evaluation while the vertical test results are included in section 6.2.

6.1. Results for Horizontal Testing.

Table 2 shows the results of the best performing agent (Agent 627) while Table 3 shows the results of the worst performing agent (Agent 612) in the Horizontal Tests.

Wins and losses are evaluated for the performance of the agents over each of the five contexts. Wins are reflected by placing the Pearson Coefficient in bold-faced blue-colored font. Coefficients that are not so highlighted represent a loss for that particular agent in the context indicated for that row. Moreover, the total win/loss tally is assisted with either a checkmark (**√**) for a win or an **x** for a loss. The number of wins for each agent over the five contexts is tallied on the last row of each agent, with the percentage of wins also indicated.

Table 4 shows the total correct output percentage over all 12 test subjects, 24 traces and five contexts. Table 5 depicts the average correct and incorrect outputs over all test subjects.

Table 6 summarizes the composite average Pearson Correlations obtained for each context over all 24 agents over the five contexts. The results roughly mirror the overall correct rates in Table 4. The relatively low standard deviations indicate consistent results over all the test subjects/agents.

Table 7 contains the results from computing the average Pearson Correlation factors for each context over the 24 agents and five contexts. This table shows that some contexts performed better than others. This is important to know, as the predictions from those contexts in which the agents performed better could be given greater confidence than those from the ones on which the agents did not perform as well. In a real-world application, the system would know the context the driver is in, and based on the cumulative history of performances under different contexts, could assign more or less weight to identifications made depending on the contexts where they were made. The agents performed best when in the Urban Construction Zone and the Urban Stop Sign contexts. We note that there is only one instance of an UCZ context in the four drives taken by the test subjects. It is possible that the networks over-trained in this context. Given the limited sample size, generalization was not necessary and the over-training yielded good results. The fact that the RCZ, which also has one instance across all four drives, also performed respectably, lends some credibility to this explanation. However, there is no context in which all agents performed well.

As can be seen from Table 8, the results for the Falconet Horizontal Tests were quite good, bordering on exceptional results for the medicated agents.

The correct rate of 81.7% for medicated and 71.7% for un-medicated were roughly validated by the Composite Average scores, which were 4/5 for each (80%).

In summary, we feel confident in declaring the Horizontal tests were successful in meeting our objectives.

Table 2
Horizontal Test – Pearson Correlation – Agent 627 (Best).

Context	Medicated Agent		p-value 95% Confidence	Win √ Loss x
	Medicated Trace	Un-medicated Trace		
RCZ	0.29	-0.05	0.0343	√
RSS	0.88	0.75	<0.0001	√
UCZ	0.54	-0.12	0.0065	√
USS	0.59	0.32	<0.0001	√
UTL	0.49	0.30	0.0449	√
	Un-medicated Agent		No. of Wins:	5/5 – 100%
	Medicated Trace	Un-medicated Trace	p-value	
RCZ	0.30	0.63	<0.0001	√
RSS	0.75	0.88	<0.0001	√
UCZ	0.41	0.49	<0.0001	√
USS	0.20	0.25	0.0156	√
UTL	0.12	0.54	0.0280	√
			No. of Wins:	5/5 – 100%

Table 3
Horizontal Test – Pearson Correlation – Agent 612 (Worst).

Context	Medicated Agent		p-value 95% Confidence	Wins √ Losses x
	Medicated Trace	Un-medicated Trace		
RCZ	0.20	0.02	<0.0001	√
RSS	0.38	0.83	<0.0001	x
UCZ	0.40	-0.05	<0.0001	√
USS	0.57	0.40	0.0268	√
UTL	0.05	0.35	0.0616	x
	Un-medicated Agent		Wins:	3/5 – 60%
	Medicated Trace	Un-medicated Trace	p-value	
RCZ	0.40	0.56	0.023	√
RSS	0.87	0.95	0.0148	√
UCZ	0.40	0.44	<0.0001	√
USS	0.50	0.27	<0.0001	x
UTL	0.39	0.21	0.0308	x
			Wins:	3/5 – 60%

Table 4
Total Wins and Losses – Horizontal Tests.

	Wins	Losses	% Wins
Medicated	49	11	**81.7%**
Un-medicated	43	17	**71.7%**

Table 5
Pearson Correlation - Average No. of Wins – Horizontal Tests.

Agent Medicated	Agent Unmedicated
4.08	3.58

6.2. *Vertical test results*

In this section, we describe the results of the Vertical tests performed. Two *Generic agents* were created – one medicated and the other un-medicated. These Generic agents were built by combining the observed human driving data on traces from test subjects #607, #608 and #613 for the medicated Generic agent, and Test Subjects #602, #608 and #620 for the un-medicated Generic agent. These subjects were chosen because they were the most similar to each other. The contexts selected for these experiments were the Rural Construction Zone (RCZ), the Rural Stop Sign (RSS) and the Urban Traffic Light (UTL). These were selected to provide a range of contexts that differed in how well the individual agents performed in them. The Generic Agent was then compared to each of the three traces that were used to train it, as well as to the other nine traces that were not. The Pearson Correlation method was used to assess similarity.

Table 9 summarizes the Vertical test results. The sub-column on the left under each context heading is the average Pearson correlation factor of the corresponding Generic Agent with each of the three human traces used to train it. The right sub-column is the average correlation with the other nine traces not used in training. Note that all comparisons are for the same medical condition (i.e., all are medicated or all are un-medicated) as this is Vertical testing and not Horizontal. The results are generally good across all three contexts measured. This suggests that

Table 6
Composite Average Pearson Correlation – Horizontal Tests.

Context	Medicated Agent				p-value 95% Confidence	Wins √ Losses x
	Medicated Trace		Un-medicated Trace			
	Mean	Std. Dev	Mean	Std. Dev		
RCZ	**0.23**	0.13	0.01	0.15	0.0294	√
RSS	0.72	0.17	0.74	0.22	0.0030	x
UCZ	**0.49**	0.10	-0.22	0.16	0.0136	√
USS	**0.58**	0.08	0.37	0.10	0.0170	√
UTL	**0.33**	0.17	0.27	0.10	0.0382	√
	Un-medicated Agent				No. of Wins:	4/5 – 80%
	Medicated Trace		Un-medicated Trace		p-value	
	Mean	Std. Dev	Mean	Std. Dev		
RCZ	0.45	0.09	**0.58**	0.08	0.0262	√
RSS	0.88	0.07	**0.93**	0.04	0.0252	√
UCZ	0.39	0.08	**0.49**	0.07	0.0122	√
USS	0.43	0.11	0.33	0.12	0.0204	x
UTL	0.36	0.13	**0.50**	0.15	0.0324	√
					No. of Wins:	4/5 – 80%

Table 7
Pearson Correlation – Win Percentage by Context – Horizontal Tests.

Contexts	Medicated Agent	Un-medicated Agent
RCZ	92%	85%
RSS	38%	77%
UCZ	100%	77%
USS	100%	38%
UTL	77%	69%

Table 8
Summary of Results for Horizontal Tests.

	SA Medicated		SA Un-Medicated	
Random	Overall	Average	Overall	Average
50%	**81.7%**	**80%**	**71.7%**	**80%**

a generic agent seems to be a viable proposition. While this clearly requires further investigation, these results are promising.

6.3. Capstone experiment - Horizontal tests with Generic agents

As a capstone experiment with the Generic agents created for the Vertical tests, it would be useful to see what differences it would make if the Horizontal tests were performed with the Generic agents created above, rather than with the individual agents being compared only to the traces of the corresponding test subject, as was done in the main Horizontal tests described in Section 6.1 above. The Generic agent now takes the place of the individual agents for each trace – that is, the same Generic agent is horizontally compared to traces of each of the 12 test subjects individually and independently. This is done for a Generic agent that represents a medicated state and another that represents an un-medicated state of the test subjects. The results are shown in Tables 10a and 10b for a medicated Generic agent and in Tables 11a and 11b for an un-medicated Generic agent. Table 10b is a continuation of 10a and Table 11b is a continuation of 11a.

Tables 10a and 10b show the correlation of the medicated Generic Agent (the same one used in the Vertical tests described above) with each of the test subject traces, both medicated and un-medicated. The correlation with the medicated traces is shown on the left half of each test subject column while the correlation with the un-medicated trace is on the right half of the test subject columns. Ideally, the correlation in the left half of the column should be higher than that on the right half of each test subject column. If such is the case, the higher left side value is highlighted in bold-face blue-colored font, indicating a win. Otherwise, there is no highlighting, indicating a loss. Note that a p-value to test for statistical significance was not calculated, so the decision of whether the comparison is a win or a loss was done on a strictly arithmetic comparison.

The total number of correct medical state identifications for the medicated Generic agent over 36 opportunities (12 test subjects x three contexts each = 36) was **28**, for a **77.8%** accuracy – significantly better than random selections, and close to the overall correctness of 81.7% in the Horizontals tests of section 6.1 for the individual medicated agents. The Generic medicated agent correctly selected **10 of 12** (**83.3%**) while in the RCZ context; **10 of 12** (**83.3%**) in RSS and **8 of 12** (**66.7%**) in the UTL context. The good performance in RCZ could be the result of over-training as we discussed above. Therefore, as indicated by the results for the individual medicated agents for each test subject, the Generic agent in this experiment can be said to have worked successfully.

Similarly, Tables 11a and 11b depict the equivalent results of Tables 10a and 10b but for the un-medicated Generic agent used in the

Table 9
Summary of Vertical Results.

Generic Agent	RCZ		RSS		UTL		Average	
Medicated	0.79	0.48	0.74	0.44	0.71	0.43	**0.75**	**0.45**
Un-medicated	0.87	0.45	0.78	0.24	0.82	0.34	**0.82**	**0.34**
Average	**0.83**	**0.47**	**0.76**	**0.34**	**0.77**	**0.39**		

Table 10a
Horizontal Test Results with Medicated Generic Agent.

	Human Test Subject Traces being Compared – Medicated/Un-medicated											
Ctxt	**602**		**607**		**608**		**609**		**612**		**613**	
RCZ	**0.44**	0.22	**0.90**	0.71	**0.45**	0.34	-0.11	-0.55	**0.11**	0.04	**0.41**	0.25
RSS	**0.55**	0.51	**0.40**	0.25	**0.43**	0.22	**0.20**	0.013	0.01	-0.55	**0.51**	0.48
UTL	-0.23	-0.15	**0.20**	0.11	**0.21**	0.16	**0.32**	0.26	**0.43**	0.32	**0.82**	0.64
Avg	0.25	0.19	0.50	0.36	0.36	0.24	0.14	-0.09	0.18	-0.06	0.58	0.46

Table 10b
Continuation of Table 10a (Medicated Generic Agent).

	Human Test Subject Traces being Compared – Medicated/Un-medicated											
Ctxt	**615**		**617**		**619**		**620**		**622**		**627**	
RCZ	**0.78**	0.42	-0.01	-0.05	**0.35**	0.31	**0.91**	-0.85	**0.81**	0.65	**0.12**	0.02
RSS	**0.33**	0.20	-0.21	-0.55	**0.67**	0.55	**0.45**	0.32	**0.77**	0.60	**0.77**	0.54
UTL	-0.34	-0.78	**0.56**	0.41	**0.76**	0.42	**0.23**	0.12	-0.11	-0.44	0.09	0.02
Avg	0.26	-0.05	0.11	-0.06	0.59	0.42	0.53	-0.13	0.49	0.27	0.33	0.19

Table 11a
Horizontal Test Results with Un-medicated Generic Agent.

	Human Test Subject Traces being Compared – Medicated/Un-medicated											
Ctxt	**602**		**607**		**608**		**609**		**612**		**613**	
RCZ	0.71	**0.81**	0.18	**0.22**	0.65	**0.77**	-0.22	-0.23	0.65	**0.66**	0.90	**0.93**
RSS	0.38	**0.43**	0.10	**0.19**	0.24	0.23	0.22	**0.32**	0.03	0.05	0.65	**0.72**
UTL	0.21	**0.27**	0.22	**0.26**	0.34	**0.54**	0.18	**0.22**	0.27	**0.31**	0.14	**0.22**
Avg	0.43	0.50	0.16	0.22	0.41	0.51	0.06	0.10	0.31	0.34	0.56	0.62

Table 11b
Continuation of Table 11a (un-medicated Generic Agent).

	Human Test Subject Traces being Compared – Medicated/Un-medicated											
Ctxt	**615**		**617**		**619**		**620**		**622**		**627**	
RCZ	-0.04	-0.05	0.44	**0.56**	0.12	**0.29**	0.91	**0.95**	-0.11	0.05	-0.066	-0.05
RSS	0.44	**0.66**	0.21	**0.23**	0.03	**0.23**	0.23	**0.38**	-0.55	-0.30	0.02	**0.12**
UTL	0.05	**0.13**	0.15	**0.27**	-0.33	-0.11	-0.05	0.05	0.14	**0.33**	0.28	**0.55**
Avg	0.15	0.25	0.27	0.35	-0.06	0.14	0.36	0.46	-0.17	0.02	0.08	0.21

Vertical tests. Ideally, the right side of each column would be higher than the left side in Tables 11a and 11b. The same highlighting in blue bold font is done in 11a and 11b to indicate correct comparisons. There is no highlighting of incorrect comparisons and no p-values were computed.

As can be gleaned from the above tables, the results for the un-medicated Generic agent are quite similar to those of the medicated Generic agent. The overall number of correct identifications was **27 of 36**, which computes to a **75.0%** rate of correctness – slightly lower than that of the medicated agent but still good, and even better than that achieved by the individual un-medicated agents of the primary Horizontal tests of Section 6.1. The context-by-context breakdown was: **8 of 12** for RCZ (**66.7%**); **9 of 12** (**75%**) for RSS; and **10 of 12** (**83.3%**) for UTL – very similar to those of the medicated Generic agent. Therefore, same conclusion of success applies to the un-medicated Generic agent.

While the overall rates of correctness for the Generic agents were near to or better than those for the individual agents, this is not an apples-to-apples comparison, as not all contexts were used in this evaluation, and the p-values were not computed. Nevertheless, their close proximity gives us cause for optimism.

6.4. Discussion of results

Overall, our experiments produced very good results. Horizontal testing resulted in 81.7% correct prediction for medicated agents and 71.7% correct prediction for un-medicated agents. Vertical testing produced 80% correct prediction for each agent using the Composite Average metric. Our data set was relatively small and we believe that a larger study would strengthen the results presented here. Nevertheless, our results suggest proof of concept that non-invasive methods could be used to monitor and detect abnormal driving behavior.

We believe that as part of further research, these numbers could be improved by making some relatively minor enhancements to Falconet, namely, further experimentation with values for its user-determined parameters. The values used here were those used by its creator, Gary Stein. Nevertheless, there are some potential threats to validity for our concept that merit discussion.

The first and most obvious threat to validity comes when making the ultimately necessary transition between a simulation and the physical world. In a simulation, everything is known or can be calculated without noise in the data. However, relying on sensors in the physical world introduces inaccuracies, noise and difficulty in interpreting the data

perceived. This issue has come to light in recent troubles with self-driving automobiles. There are two parts to this threat. The first is that the agents will not be able to perceive the physical environment as well as the human, and therefore introduce a divergence that may be unwarranted. However, more importantly, if the traces to be used for training are to come from actual driving experience, then collecting those data may be difficult, given the cost to instrument automobiles. We mentioned earlier that a sensor suite may become less costly as self-driving cars become more accepted, but that could be considered a bold prediction at this time. Of course, one could build the agents from traces gathered in simulators, as we did here, but how representative the simulated drives are of actual driving experiences is still a question. It is true that simulators have been extensively and effectively used in training for aircraft pilots since the 1960 s; however, they are always accompanied by live training in actual airplanes.

Another threat is how an agent would react when facing a situation that has never been faced before. This is also a problem in self-driving automobiles. The machine learning process used must be able to show a non-trivial level of generalization ability in order to successfully manage such contexts. In fact, testing for and improving generalization is one of our main objectives in future research. Stein found that Falconet displayed a good ability to generalize, but that needs to be shown in this type of application.

7. Summary, conclusions and future research

To summarize, we investigated an approach based on the use of Machine Learning from Observation to discriminate between an ADHD-afflicted driver driving when medicated or when un-medicated. The approach involved a tool for LfO called Falconet that used Neuro-evolution, Particle Swarm Optimization and Context-based Reasoning to build agents that reflect the driving behavior of the observed human.

The resulting agents were created from observation of 12 human test subjects driving automobiles in simulations. By comparing the actions predicted with the actual actions taken by the driver, we can infer his/her medication state. In the future, these models can be installed on the automobiles that afflicted drivers would drive in the real world to provide a real time identification of dangerous driving conditions on the part of the driver. This can make a transformational improvement in road safety without intrusive physical monitoring of the driver.

In conclusion, we believe that the major outcome of our work has been that our results verify our hypothesis stated in Section 2 above and re-printed below for the benefit of the reader.

> Models of human driving behavior, built through machine learning from observation of human drivers afflicted with ADHD, could be compared to their actual driving behavior to identify when these drivers are operating a motor vehicle while suffering from un-controlled ADHD conditions.

Driving models built with the Falconet architecture seem to be capable of detecting the nuanced differences between medicated and un-medicated behaviors. Our Horizontal test results strongly suggest the scientific feasibility of using a model of driving behavior of individuals who are afflicted with ADHD to detect whether they are driving while in an un-medicated condition. However, we stop short of asserting that our work here is proof of such feasibility, as the sample size used was too small for such an assertion. Moreover, we believe that our experiments should be expanded, as we discuss below. We hope to undertake those tests in future research.

The Vertical tests suggest that the concept of building generic agents can work. This would be a significant advantage over the need to create individual agents for each person to be monitored in the context of a commercial application of this concept. Nevertheless, we do not believe that the latter is necessarily a showstopper if this technique were to be applied to high risk drivers, whose health is such that individualized modeling of his/her behavior is warranted.

Our results were achieved with a relatively small amount of data. In this era when Big Data is the in-fashion application, the ability to use "small data" could be advantageous, as positive outcomes could be achieved in situations where collecting these data is expensive and/or difficult to obtain, as was certainly the case in our work. Recruiting test subjects was not easy and the test subjects that were recruited had to be handled carefully and thoughtfully because of the sensitive nature of the data we sought to collect.

Nevertheless, our work is a long way away from technical and commercial feasibility. There were several assessments that we did not perform as a result of lack of resources, lack of data, lack of time or all of the above. The first one to come to mind is the acquisition and use of context transition rules. In our context-centric approach, identification of the context being faced by the driver is essential, and it must be done correctly. We regrettably must leave that for future research.

Also regrettably left for future research is determining how to assess the generality of the agents being built. In other words, do they learn to handle all stop signs? Or just the ones used in training? A measure of this was implicit in our testing, as there were several instances of stop signs and traffic lights, as well as in the concept of a generic agent, but it would be beneficial to do this explicitly and more rigorously.

Another item for future research is to automate the decomposition of the traces into context instances for training the agents. This would facilitate the contextualization of observational data (traces) and make it easier to train the agents. Trinh (Trinh & Gonzalez, 2013) addressed this problem and developed a tool called COPAC (COntext Partitioning And Clustering) as part of his doctoral dissertation, and obtained good results. However, we chose to not use it in order to maintain the focus of our limited resources on assessing the scientific feasibility of the overall concept.

Other issues still to be addressed include how to incorporate this concept into a real-time tool used in the physical world, with the assorted difficulties such an undertaking brings. Certainly the thought of creating test cars with the appropriate sensors comes to mind first. The instrumentation required in the car would be our second thought to arise. Even beyond that, however, would be how to continuously compare the output of the agent with the actions being performed by the driver in real time, and when is there enough evidence of serious discrepancies to yell out "Bingo", so to speak.

One final area of further research involves how well will agents built with data from simulations work when placed in the physical world. While data to train agents can be obtained from driving an actual automobile in real traffic, it could only be done when the human subject is medicated, as it would present too much risk to have an un-medicated driver purposely driving public roads.

CRediT authorship contribution statement

Avelino J. Gonzalez: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision. **Josiah M. Wong:** Software, Investigation, Formal analysis. **Emily M. Thomas:** Software, Investigation. **Alec Kerrigan:** Software, Investigation. **Lauren Hastings:** Software, Investigation. **Andres Posadas:** Software, Investigation. **Kevin Negy:** Software, Investigation. **Annie S. Wu:** Investigation, Methodology. **Santiago Ontañon:** Methodology, Funding acquisition. **Yi-Ching Lee:** Data curation, Funding acquisition. **Flaura K. Winston:** Conceptualization, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported by the U.S. National Science Foundation grant No. SCH-1521972, under its Smart, Connected Health Program - Dr. Wendy Nilsen, Program Director.

References

Aduen, P. A., Cox, D. J., Fabiano, G. A., Garner, A. A., & Kofler, M. J. (2019). Expert recommendations for improving driving safety for teens and adult drivers with ADHD. *The ADHD Report, 27*(4), 8–14.

Aihe, D. O., & Gonzalez, A. J. (2015). Correcting flawed expert knowledge through reinforcement learning. *Expert Systems with Applications., 42*(17–18), 6457–6471.

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems., 57*(5), 469–483.

Barkley, R. A., Guevremont, D. C., Anastopoulos, A. D., DuPaul, G. J., & Shelton, T. L. (1993). Driving-related risks and outcomes of attention deficit hyperactivity disorder in adolescents and young adults: A 3-to 5-year follow-up survey. *Pediatrics., 92*(2), 212–218.

Barkley, R. A., Murphy, K. R., O'Connell, T., & Connor, D. F. (2005). Effects of two doses of methylphenidate on simulator driving performance in adults with attention deficit hyperactivity disorder. *Journal of Safety Research., 36*(2), 121–131.

Biederman, J., Fried, R., Hammerness, P., Surman, C., Mehler, B., Petty, C. R., … Reimer, B. (2012). The effects of lisdexamfetamine dimesylate on driving behaviors in young adults with ADHD assessed with the Manchester Driving Behavior Questionnaire. *Journal of Adolescent Health., 51*(6), 601–607.

Boland, H., DiSalvo, M., Fried, R., Woodworth, K. Y., Wilens, T., Faraone, S. V., & Biederman, J. (2020). A literature review and meta-analysis on the effects of ADHD medications on functional outcomes. *Journal of Psychiatric Research., 123*, 21–30.

Curry, A. E., Metzger, K. B., Pfeiffer, M. R., Elliott, M. R., Winston, F. K., & Power, T. J. (2017). Motor vehicle crash risk among adolescents and young adults with Attention-Deficit/Hyperactivity Disorder. *JAMA Pediatrics, 171*(8), 756–763.

Curry, A. E., Yerys, B. E., Metzger, K. B., Carey, M. E., & Power, T. J. (2019). Traffic crashes, violations, and suspensions among young drivers with ADHD. *Pediatrics, 143*(6). https://doi.org/10.1542/peds.2018-2305

Das, D., Zhou, S., & Lee, J. D. (2012). Differentiating alcohol-induced driving behavior using steering wheel signals. *IEEE Transactions on Intelligent Transportation Systems, 13*(3), 1355–1368.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4–7.

Fabiano, G. A., Schatz, N. K., Morris, K. L., Willoughby, M. T., Vujnovic, R. K., Hulme, K. F., … Pelham, W. E., Jr. (2016). Efficacy of a family-focused intervention for young drivers with Attention-deficit Hyperactivity Disorder. *Journal of Consulting and Clinical Psychology, 84*(12), 1078.

Faraoner, S. V., Rostain, A. L., Blader, J., Busch, B., Childress, A. C., Connor, D. F., & Newcorn, J. H. (2019). Practitioner Review: Emotional dysregulation in attention-deficit/hyperactivity disorder – implications for clinical recognition and intervention. *Journal of Child Psychology and Psychiatry, 60*(2), 133–150.

Fernlund, Hans. Evolving Models from Observed Human Performance. Dissertation, University of Central Florida. 2006. pg. 59, 122, 128.

Gonzalez, A. J., Stensrud, B. S., & Barrett, G. (2008). Formalizing Context-Based Reasoning - A modeling paradigm for representing tactical human behavior. *International Journal of Intelligent Systems, 23*(7), 822–847.

Grethlein, D. and Ontañón, S. (2020). *Spatially aligned clustering of driving simulator data.* Proceedings of the Florida Artificial Intelligence Research Society Conference (FLAIRS) 2020.

Grethlein, D., Winston, F. K., Walshe, E., Tanner, S., Kandadai, V., & Ontañón, S. (2020). Simulator pre-screening of underprepared drivers prior to licensing on-road examination: Clustering of virtual driving test time series data. *Journal of Medical Internet Research, 22*(6), Article e13995. https://doi.org/10.2196/13995

Groom, M. J., Cahill, J. D., Bates, A. T., Jackson, G. M., Calton, T. G., Liddle, P. F., & Hollis, C. (2010). Electrophysiological indices of abnormal error-processing in adolescents with attention deficit hyperactivity disorder (ADHD). *Journal of Child Psychology and Psychiatry, 51*(1), 66–76.

Groom, M. J., van Loon, E., Daley, D., Chapman, P., & Hollis, C. (2015). Driving behaviour in adults with attention deficit/hyperactivity disorder. *BMC Psychiatry, 15*, 175.

Hollister, D.L., Gonzalez, A.J. & Hollister, J.R. (2019). Contextual Reasoning in Human Cognition and its Implications for Artificial Intelligence Systems. *Modelisation et Utilisation du Contexte.* iSTE OpenScience. June 2019. 1-18. https://www.openscience.fr/IMG/pdf/iste_muc19v3n1_1.pdf.

Jin, L., Niu, Q., Jiang, Y., Xian, H., Qin, Y., & Xu, M. (2013). Driver sleepiness detection system based on eye movements variables. *Advances in Mechanical Engineering, 5*, Article 648431.

Johnson, C. L., & Gonzalez, A. J. (2014). Learning collaborative behavior by observation. *Expert Systems with Applications, 41*, 2316–2328.

Kang, H.-B. (2013). Various approaches for driver and driving behavior monitoring: A Review. *IEEE International Conference on Computer Vision Workshops (ICCVW)*.

Karatekin, C. (2007). Eye tracking studies of normative and atypical development. *Developmental Review, 27*(3), 283–348.

Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks, 4*, 1942–1948.

Kishimoto, Y., Abe, K., Miyatake, H., & Oguri, K. (2008). Modeling driving behavior with dynamic bayesian networks and estimate of mental state. *15th World Congress on Intelligent Transport Systems and ITS America's 2008 Annual Meeting*.

Koza, J. R. (1992). *Genetic Programming.* Cambridge MA: MIT Press.

Lee, Y.-C., Ward McIntosh, C., Winston, F., Power, T., Huang, P., Ontañón, S., & Gonzalez, A. J. (2018). Design of an experimental protocol to examine medication non-adherence among young drivers diagnosed with ADHD: A driving simulator study. *Contemporary Clinical Trials Communications, 11*, 149–155.

Li, Q., Zhao, L., Lee, Y-C. and Lin, J. (2020a). Contrast pattern mining in paired multivariate time series of a controlled driving behavior experiment. *ACM Transactions Spatial Algorithms Syst.* 6(4), Article 25 28 pages. doi:10.1145/3397272.

Li, Q., Zhao, L., Lee, Y.-C., Sassanin, A., & Lin, J. (2020). CPM: A general feature dependency pattern mining framework for contrast multivariate time series. *Pattern Recognition, 107711*. https://doi.org/10.1016/j.patcog.2020

Liang, Y., (2009) Detecting driver distraction. *Theses and Dissertations*, 248.

Merkel, R. L., Jr., Nichols, J. Q., Fellers, J. C., Hidalgo, P., Martinez, L. A., Putziger, I., … Cox, D. J. (2016). Comparison of on-road driving between young adults with and without ADHD. *Journal of Attention Disorders, 20*(3), 260–269.

Oliver, N., & Pentland, A. P. (2000). Graphical models for driver behavior recognition in a smartcar. *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*.

Otmani, S., Pebayle, T., Roge, J., & Muzet, A. (2005). Effect of driving duration and partial sleep deprivation on subsequent alertness and performance of car drivers. *Physiology & Behavior, 84*(5), 715–724.

Randell, N. J. S., Charlton, S. G., & Starkey, N. J. (2020). Driving with ADHD: Performance effects and environment demand in traffic. *Journal of Attention Disorders, 24*(11), 1570–1580.

Reimer, B., Mehler, B., D'Ambrosio, L. A., & Fried, R. (2010). The impact of distractions on young adult drivers with attention deficit hyperactivity disorder (ADHD). *Accident Analysis & Prevention, 42*(3), 842–851.

Sahayadhas, A., Sundaraj, K., & Murugappan, M. (2012). Detecting driver drowsiness based on sensors: A review. *Sensors, 12*(12), 16937–16953.

Salvucci, D. D. (2004). Inferring driver intent: A case study in lane-change detection. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*.

Sidani, T. A., & Gonzalez, A. J. (2000). A framework for learning implicit expert knowledge through observation. *Transactions of the Society for Computer Simulation, 17*(2), 54–72.

Siordia, O. S., de Diego, I. M., Conde, C., Reyes, G., & Cabello, E. (2010). Driving risk classification based on experts evaluation. *Intelligent Vehicles Symposium (IV)*.

Stanley, K., & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation, 10*(2), 99–127.

Stein, G., & Gonzalez, A. J. (2011). Building high-performing human-like tactical agents through observation and experience. *IEEE Transactions on Systems, Man and Cybernetics - Part B, 41*(3), 792–804.

Stein, G., & Gonzalez, A. J. (2014). Learning in Context: Enhancing Machine Learning with Context-Based Reasoning. *Applied Intelligence, 41*, 709–724.

Stein, G., Gonzalez, A. J., & Barham, C. (2015). Combining NEAT and PSO for learning tactical human behavior. *Neural Computing and Applications, 26*(4), 747–764.

Stein, G. (2009). *Falconet: force-feedback approach for learning from coaching and observation using natural and experiential training.* Doctoral Dissertation. Computer Engineering, University of Central Florida, August 2009.

Stensrud, B. S., & Gonzalez, A. J. (2008). Discovery of high-level behavior from observation of human performance in a strategic game. *IEEE Transactions on Systems, Man and Cybernetics, Part B, 38*(3), 855–874.

Suzuki, T., Sekizawa, S., Inagaki, S., Hayakawa, S., Tsuchida, N., Tsuda, T., & Fujinami, H. (2005). Modeling and recognition of human driving behavior based on stochastic switched ARX model. *44th IEEE Conference on.Decision and Control and European Control Conference CDC-ECC'05*.

Torabi, F., Warnell, G. & Stone, P. (2019). Recent advances in imitation learning from observation. *arXiv preprint arXiv:1905.13566*.

Torkkola, K., Massey, N., & Wood, C. (2008). Driver inattention detection through intelligent analysis of readily available sensors. *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems*.

Trinh, V. C., & Gonzalez, A. J. (2013). Identifying contexts from observed human performance. *IEEE Transactions on Human Machine Systems, 43*(4), 359–370.

Vaa, T. (2014). ADHD and relative risk of accidents in road traffic: A meta-analysis. *Accident Analysis & Prevention, 62*, 415–425.

Wong, J. M., Hastings, L., Negy, K. Gonzalez, A.J., Ontañón, S., and Lee, Y-C. (2018). *Machine learning from observation to detect abnormal driving behavior in humans*. Proceedings of the 31st Annual Florida Artificial Intelligence Research Society Conference (FLAIRS-2018). May 2018.

Zibetti, E., Hamilton, E., & Tijus, C. (1999). The role of context in interpreting perceived events as actions. *International and Interdisciplinary Conference on Modeling and Using Context*, 431–441.