Serious Gaming for Building a Basis of Certification via Trust and Trustworthiness of Autonomous Systems

B. Danette Allen[1]
NASA Langley Research Center, Hampton, VA, 23681, *USA*

Autonomous systems governed by a variety of adaptive and nondeterministic algorithms are being planned for inclusion into safety-critical environments, such as unmanned aircraft and space systems in both civilian and military applications. However, until autonomous systems are proven and perceived to be capable and resilient in the face of unanticipated conditions, humans will be reluctant or unable to delegate authority, remaining in control aided by machine-based information and decision support. Proving capability, or trustworthiness, is a necessary component of certification. Perceived capability is a component of trust. Trustworthiness is an attribute of a cyber-physical system that requires context-driven metrics to prove and certify. Trust is an attribute of the agents participating in the system and is gained over time and multiple interactions through trustworthy behavior and transparency. Historically, artificial intelligence and machine learning systems provide answers without explanation – without a rationale or insight into the machine "thinking". In order to function as trusted teammates, machines must be able to explain their decisions and actions. This transparency is a product of both content and communication. NASA's Autonomy Teaming & TRAjectories for Complex Trusted Operational Reliability (ATTRACTOR) project seeks to build a basis for certification of autonomous systems via establishing metrics for trustworthiness and trust in multi-agent team interactions, using AI explainability and persistent modeling and simulation, in the context of mission planning and execution, with analyzable trajectories. Inspired by Massively Multiplayer Online Role Playing Games (MMORPG) and Serious Gaming, the proposed ATTRACTOR modeling and simulation environment is similar to online gaming environments in which player (aka agent) participants interact with each other, affect their environment, and expect the simulation to persist and change regardless of any individual agent's active participation. This persistent simulation environment will accommodate individual agents, groups of self-organizing agents, and large-scale infrastructure behavior. The effects of the emerging adaptation and co-evolution can be observed and measured to building a basis of measurable trustworthiness and trust, toward certification of safety-critical autonomous systems.

I. Nomenclature

a priori	=	from the earlier
a posteriori	=	from the latter
ab initio	=	from the beginning
AFRC	=	Armstrong Flight Research Center
DARPA	=	Defense Advanced Research Projects Agency
AI	=	Artificial Intelligence
XAI	=	eXplainable AI
ATTRACTOR	=	Autonomy Teaming & TRAjectories for Complex Trusted Operational Reliability
LaRC	=	Langley Research Center
NASA	=	National Aeronautics and Space Administration
MMORPG	=	Massively Multiplayer Online Role Playing Game

[1] NASA Senior Technologist for Intelligent Flight Systems, NASA LaRC, MS 492, AIAA Senior Member.

II. Introduction

As we look towards a future where autonomous systems are a ubiquitous part of our lives, we must overcome significant technical challenges to ensure reliable and safe operations. Be it in our homes, on our roads, or in our airspace, an autonomous system functioning in safety-critical and/or time-critical environments must be verified and validated to be certified safe. When these systems are mission critical, human safety may not be a concern but a successful outcome may be entirely dependent on a system's ability to manage itself and its environment. When these systems are truly autonomous (as opposed to automated) and are

(1) expected to achieve complex goals while operating independently of external control

(2) often nondeterministic and/or adaptive due to stochastic methods and machine learning capabilities

our historical assumptions about determinism and behavior boundaries no longer apply.

Historically, systems have been proven reliable and safe via rigorous and thorough methods [1] while strenuous system and software engineering processes [2, 3, 4] helped to ensure responsible development. Operational assurance was achieved through extensive testing of all possible deterministic paths. The shift away from the "if-then" paradigm that underpins these methods demands new approaches for certifying systems. One possible approach is a flight simulation infrastructure for *ab initio* (or clean slate) modeling and simulation that assumes no specific architecture and models agent-to-agent behavior to examine interactions and emergent behaviors among potentially hundreds or thousands of intelligent agents exhibiting myriad behaviors.

III. ATTRACTOR

ATTRACTOR (Autonomy Teaming & TRAjectories for Complex Trusted Operational Reliability) is a new research effort started at NASA under the Transformative Aeronautics Concepts Program (TACP) Convergent Aeronautics Solutions (CAS) Project focused on building a basis of certification for autonomous systems via defining metrics for trust and trustworthiness and building a simulation environment in which these concepts can be explored. Proving capability, or trustworthiness, is a necessary component of certification. Perceived capability is a component of trust. Trust is an attribute of the agents participating in the system and is gained over time and multiple interactions through trustworthy behavior and transparency. Trustworthiness is an attribute of a cyber-physical system that requires context-driven metrics to prove and certify. To build a basis of certification for autonomous systems, attention must be paid to both content (for trustworthiness) and communication (for trust).

Today's artificial intelligence and machine learning systems provide answers without explanation – without a rationale or insight into the machine "thinking". In order to function as trusted teammates with humans and other agents, machines must be able to explain their decisions and actions. This transparency in decision-making (whether in real-time or otherwise) is a critical aspect of the ATTRACTOR research portfolio. Explainable Artificial Intelligence (XAI) is an effort to make intelligent machine systems more understandable and is especially relevant to applications that rely on machine learning and/or nondeterministic algorithms. This explainability is key to effective integration of these systems and NASA is not alone in our need to increase the transparency of these systems. For example, DARPA's XAI Program [5] is striving to produce more explainable models while maintaining a high level of learning performance while NASA is focused on explainability in the training and operational context.

The goal of ATTRACTOR is make progress towards building a basis for certification of autonomous systems

- via establishing metrics for trustworthiness and trust
- in single- and multi-agent team interactions
- using XAI and
- and analyzable trajectories

in the operational context of mission-critical planning and execution. A persistent modeling and simulation environment for a-priori, real-time and a-posteriori interaction for test and evaluation is the foundation upon which these goals rest. This simulation environment will accommodate individual agents, groups of self-organizing agents, and large-scale infrastructure behavior. Inspired by Massively Multiplayer Online Role Playing Games (MMORPG) and Serious Gaming, the ATTRACTOR simulation environment is similar to online gaming environments in which player participants interact with each other, affect their environment, and expect the simulation to persist and change regardless of any individual agent's active participation.

IV. Serious Gaming

A *serious game* is a game designed specifically for the purpose of solving real problems such as training, learning, etc. and not primarily for the purpose of entertainment. The use of serious games to gain insight into societal challenges, military scenarios, air traffic management [6], emergency operations, etc. has shown potential over the last decade. In 2007, the UK experienced devastating floods (Figure 1) across the country. The total economic cost [7] was estimated to be about £3.2 billion.

Figure 1: London Victoria Station Flooded in 2007 [8]

In response, the FloodSim™ serious game simulation (Figure 2) was released in 2008 to raise public awareness of the risk and hazards of flooding in the UK by asking public participants to set and execute policy over a simulated three-year period. Over only four weeks in August/September 20118, over 25,000 players participated in the FloodSim™ simulation.

Figure 2: A simulation of floods in London and the FloodSim™ Interface [9]

Studies [9] have suggested that, while this serious gaming effort connected with citizens around a government policy mechanism that they would likely not have otherwise engaged, determining whether there was an increased awareness around flood issues proved difficult. However, it is the number of participants and number of simulation hours enabled by serious gaming that shows promise for autonomous systems, especially systems that rely on machine learning.

V. Persistent Simulation

Machine learning applications can be data-hungry and gaining sufficient access to hard-to-get training data appropriate for autonomous intelligent flight systems such as trajectories for mission planning and specialized images for computer vision classification (among others) has proven to be a challenge. Building on the existing AEON (Autonomous Entity Operations Network) framework [10], a high-fidelity serious-gaming environment could support the generation of hundreds or thousands of hours, events, images, video, etc. for use by intelligent agents seeking to learn from as they “play” the serious game or, more formally, the serious persistent simulation.

A persistent simulation is a virtual world [11] that "continues to exist and develop internally even when there are no people interacting with it". Typically, this refers to MMORPGs but also relates to “pervasive games” in which the real and virtual worlds are blended. This blending is not unlike the Live Virtual Constructive environments in use today in NASAs Air Traffic Management (ATM) simulation facilities across the country in distributed locations to participate in a shared simulation environment. However, these environments were designed with specific concepts of operations (ConOps) in mind so are not pervasive and are not designed for exploring a solution space that includes innovative architectures, heterogeneous vehicles, changing “rules of the road”, and player participants that enter and exit the simulation with their solutions at will.

Games such as World of Warcraft ™ and Minecraft ™ (for example) are essentially distributed simulations inside which avatars controlled by humans interact with their environment and other avatars (aka participants or agents) in a world (or a level) of choice and make decisions that have consequences within these MMORPGs. If we repurpose the characteristics and capabilities of MMORPGs along with the autonomy-enabling technologies to support persistent simulation of autonomous systems, we can test, evaluate, and observe the behavior of the human and/or machine participants as they pursue individual or teamed mission goals.

For ATTRACTOR, we are developing Baseline Environment for Autonomous Modeling (BEAM) to serve as a software testbed for autonomous capability development and to support our exploration of XAI. Built on the Unity™ game engine and utilizing Data Distribution Service (DDS) middleware, BEAM utilizes a “persistence server” [12] that is a hybrid between a traditional server and a peer-to-peer (P2P) server to monitor agent actions and record the world state. Players can login (Figure 3) as an “observer” or player type.

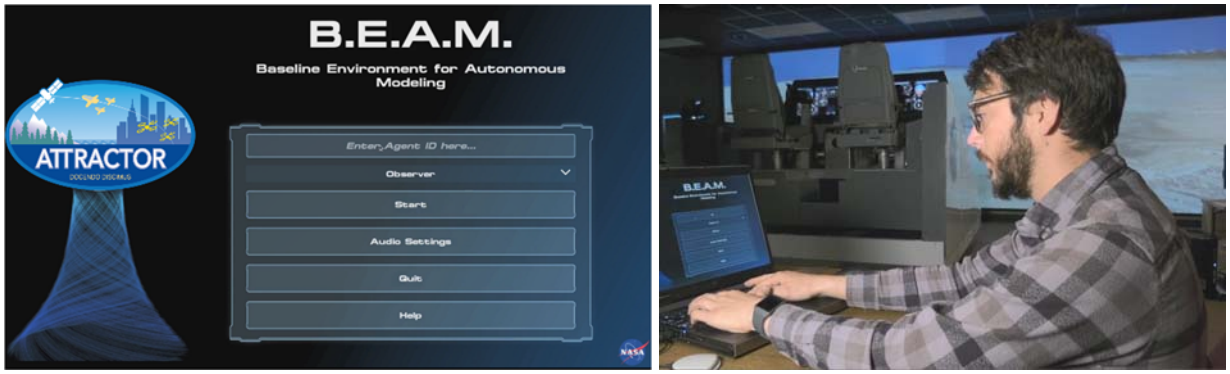

Figure 3: ATTRACTOR BEAM Simulation Login

Player types can be humanoid, ground-based, air-based, etc. and will function according to the physics-based rules of the game engine and simulation world or level. Observers can gather data from the players (agents) for use in test and evaluation but also for use as training data. For example, if an aerial agent is equipped with a vision system such as a lidar or camera, simulated sensor data is available over DDS for use in playback or in real-time. As shown in Figure 4, BEAM is already capable of fusing real and virtual assets and an immersive functionality for use in planning and execution is in development. The infrastructure is in place for distributed simulation and participants at LaRC and Armstrong Flight Research Center (AFRC) have demonstrated proof-of-concept with two agents from LaRC and one from AFRC interacting on the “LaRC AI” level as three independent ground vehicles as shown in Figure 5.

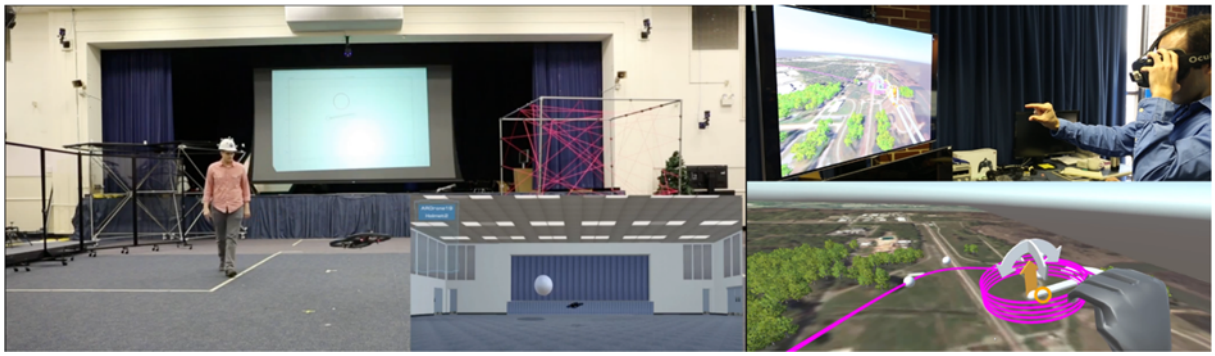

Figure 4: ATTRACTOR BEAM Simulation Mixed Reality Environment

Figure 5: ATTRACTOR BEAM Simulation Distributed Environment

VI. Conclusion and Future Work

The persistent simulation created for ATTRACTOR is a work in progress. BEAM is already capable of creating and maintaining mixed reality worlds in real-time with distributed agents. A playback capability has been created and is being used for offline test and evaluation of recent flight testing in the Autonomy Incubator facility at LaRC. The AI community's excitement about the recent release of the "StarCraft AI Research Dataset" for "a wide variety of machine learning tasks such as strategy classification, inverse reinforcement learning, [and] imitation learning" [13] is an encouraging sign that this playback function will be valuable. Intuitive human-machine interfaces (HMI) utilizing natural language and gestures in support of trust [14] are being evaluated and incorporated into BEAM. Large scale simulation worlds that encompass all participating NASA centers will demand a shift away from the naïve localized map datum to a geodetic coordinate system that accounts for the ellipsoid shape of the earth for coast-to-coast compatibility. Very soon, ATTRACTOR researchers will be generating their own "mods" to the BEAM simulation environment to create agents running their trajectory-generation or object-classification algorithms for test in a realistic virtual environment. Simulations, whether serious games or not, do not eliminate the need for actual flight testing. Flight testing is time-consuming and expensive so we are hopeful that serious gaming simulations such as

BEAM will facilitate an agile software-in-the-loop and perhaps hardware-in-the loop test and evaluation environment that will accelerate training, testing, and deployment of autonomous intelligent systems.

Acknowledgments

This work would not have been possible without the support of NASA LaRC and, in particular, Pete Lillehei for seeing the potential in the ATTRACTOR idea while still part of the Autonomy Incubator. This sponsorship accelerated our thinking and development, leading to ATTRACTOR's selection for a 2018 start in NASA ARMD's TAC CAS project. Feasibility of the idea could not have been demonstrated without the talented software development team from AMA (Ben Kelley, Kyle McQuarry, Matt Vaughan, and Ralph Williams) and the contribution of Autonomy Incubator student interns (Meghan Chandarana, Erica Meszaros, and Angelica Garcia). Last but not least, many thanks to Natalia Alexandrov for her words in this abstract and for being a remarkable collaborator and co-PI.

References

[1] Christine M. Belcastro, "Validation and Verification (V&V) of Safety-Critical Systems Operating under Off-Nominal Conditions", 2018-05-31, https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20140003450.pdf.

[2] NASA Systems Engineering Processes and Requirements, NPR 7120.8, https://nodis3.gsfc.nasa.gov/displayDir.cfm?t=NPR&c=7120&s=8

[3] NASA Systems Engineering Processes and Requirements, NPR 7123.1B, https://nodis3.gsfc.nasa.gov/displayDir.cfm?t=NPR&c=7123&s=1B.

[4] NASA Systems Engineering Handbook (SP-2016-6105), Rev 2, http://hdl.handle.net/2060/20170001761.

[5] Explainable Artificial Intelligence (XAI), https://www.darpa.mil/program/explainable-artificial-intelligence

[6] Bonnie D. Allen and Natalia Alexandrov. "Serious Gaming for Test & Evaluation of Clean-Slate (Ab Initio) National Airspace System (NAS) Designs", 16th AIAA Aviation Technology, Integration, and Operations Conference, AIAA AVIATION Forum, (AIAA 2016-4375), https://doi.org/10.2514/6.2016-4375

[7] "The costs of the summer 2007 floods in England", Project: SC070039/R1, The Environment Agency/Defra Flood and Coastal Erosion Risk Management Research and Development Programme, ISBN: 978-1-84911-146-1, January 2010.

[8] Frankie Roberto [Public domain], Wikimedia Commons, https://commons.wikimedia.org/wiki/File:London_Victoria_Station_flooded.jpg

[9] Rebolledo-Mendez, Genaro & Avramides, Katerina & de Freitas, Sara & Star, Kam. (2009). Societal impact of a serious game on raising public awareness: The case of FloodSim. Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games, Sandbox '09. 15-22. 10.1145/1581073.1581076.

[10] Charles D. Cross, Mark A. Motter, James H. Neilan, Garry D. Qualls, Paul M. Rothhaar, Loc Tran, Anna C. Trujillo, and B. Danette Allen "Towards an Open, Distributed Software Architecture for UxS Operations" 2016.

[11] Richard Bartle, "Designing Virtual Worlds", New Riders, ISBN 0-13-101816-7, 2003.

[12] Benjamin N. Kelley, Jason L. Holland, Otto C. Schnarr, Ralph A. Williams, and B. Danette Allen, "A Persistent Simulation Environment for Autonomous Systems", to be published at 18th AIAA Aviation Technology, Integration, and Operations Conference, AIAA AVIATION Forum, Atlanta, GA, 2018.

[13] https://github.com/TorchCraft/StarData

[14] Meghan Chandarana, Erica L. Meszaros, Anna Trujillo, and B. Danette Allen, "Natural Language Based Multimodal Interface for UAV Mission Planning", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol 61, Issue 1, pp. 68 – 72, September 2017.