3-2022

Obsolescence: Evaluating an Educational Serious Game on Artificial Intelligence Impacts to Military Strategic Goals

Timothy C. Kokotajlo

‘OBSOLESCENCE:’ EVALUATING AN EDUCATIONAL SERIOUS GAME ON ARTIFICIAL INTELLIGENCE IMPACTS TO MILITARY STRATEGIC GOALS

THESIS

Timothy C Kokotajlo, Captain, USAF

AFIT-ENG-MS-22-M-039

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-22-M-039

‘OBSOLESCENCE:’ EVALUATING AN EDUCATIONAL SERIOUS GAME ON ARTIFICIAL INTELLIGENCE IMPACTS TO MILITARY STRATEGIC GOALS

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cyber Operations

Timothy C Kokotajlo, BS

Captain, USAF

March 2022

AFIT-ENG-MS-22-M-039

‘OBSOLESCENCE:’ EVALUATING AN EDUCATIONAL SERIOUS GAME ON ARTIFICIAL INTELLIGENCE IMPACTS TO MILITARY STRATEGIC GOALS

Timothy C Kokotajlo, BS
Captain, USAF

Committee Membership:

Dr. Mark Reith
Chair

Dr. David Long
Member

Dr. Gilbert Peterson
Member

AFIT-ENG-MS-22-M-039

Abstract

Artificial Intelligence (AI) threatens to bring significant disruption to all aspects of military operations. This research develops a Serious Game (SG) and assessment methodology to provide education on the mindsets required for engaging with disruptive AI technologies. The game, Obsolescence, teaches strategic-level concepts recommended to the Department of Defense (DoD) from a compilation of reports on the current and future state of AI and warfighting. The methodology for assessing the educational value of Obsolescence addresses common challenges such as subjective reporting, control groups, population sizes, and measuring abstract or high levels of learning. The game's proposed educational value is tested using a pre- and post-test format against a baseline established by official sources and experts in the fields of AI and strategic planning. The assessment includes metrics based on both self-reported learning and measurements of changes to participant responses to LO-related questions post-gameplay. The experiment found a strong correlation between the measured learning and participants' self-reported learning, and both metrics confirm that Obsolescence achieves its educational goals. This research includes the steps necessary to utilize the assessment methodology and presents recommendations both for Obsolescence and for future research in the field of educational game assessment.

Table of Contents

Page

Abstract iv
List of Figures viii
List of Tables ix
I. Introduction 1
Purpose and Problem Statement 1
Background and Motivation 2
Research Questions 3
Approach 4
Scope and Limitations 6
Contributions 8
Chapter Structure 8
II. Background 10
Chapter Overview 10
Serious Games (SGs) 10
Disruptive Artificial Intelligence Technologies 20
Tools 24
Background Summary 25
III. Design of Environment 26
Chapter Overview 26
Obsolescence's Learning Objectives (LOs) 27

Design of Game Mechanics....36
Relations to Educational Goals....46
Relations to Research Questions....49
Conclusion....51
IV. Methodology....52
Chapter Overview....52
Game Assessment Methodology....52
Experiment Design for Obsolescence....62
Conclusion....67
V. Results and Analysis....69
Chapter Overview....69
Data Preparation....69
Establishing the Baseline....70
Reported Learning....72
Measured Learning....73
Comparison Between Reported and Measured Learning....77
Engagement....80
Conclusion and Other Notes....85
VI. Conclusion....87
Chapter Overview....87
Research Summary....87
Research Contributions....90
Observations on Research Procedures and Lessons Learned....93
Future Work....95

Final Thoughts....98

Appendix A: Obsolescence Rules....99

Appendix B: Survey Questions....107

Appendix C: Additional Data Tables and Figures....110

VII. Bibliography....113

List of Figures

Page

Figure 1: Bloom's Affective Taxonomy [10] 12

Figure 2: Bloom's Cognitive Taxonomy [21] 13

Figure 3: Techniques used to evaluate SGs, in number and % of 102 papers [22] 17

Figure 4: Common Questionnaire types [22] 18

Figure 5: Distribution of SG study population sizes, displayed both between 1->120 participants and 1-40 participants [22] 19

Figure 6: Sample Obsolescence screenshot with labeled interface icons. 38

Figure 7: The AI's decision tree for each turn's actions 44

Figure 8: Methodology handout describing the assessment's procedural flow 53

Figure 9: Reported Learning for each LO 72

Figure 10: Average change towards baseline per LO 74

Figure 11: Average change in participant answers relative to the baseline, per question. Higher values indicate a stronger change towards the baseline. 75

Figure 12: Graphical representation between measured and reported Learning, normalized, per participant 79

Figure 13: Responses to Post Survey Part Two-Question One 81

Figure 14: Responses to Post Survey Part Two-Question Two 82

Figure 15: Comparison of reported learning and Part Two of Post-Survey. Higher values in one dimension correlate with higher values in other dimensions 83

List of Tables

Page

Table 1: List of Learning Objectives (LOs) for Obsolescence, alongside their intended Cognitive (C) and Affective (A) Taxonomy level and source documents..... 28

Table 2: Comparison of Obsolescence and Hedgemony[11] 37

Table 3: Tech Card effects, including related LOs, how early they can appear in the game, their cost, the game effect, and the flavor text displayed to players..... 42

Table 4: Possible Inputs and outputs to AI decision-making 45

Table 5: Final weight set for Obsolescence's in-game AI opponents 45

Table 6: The organizations given access to Obsolescence 63

Table 7: Baseline generation using experts and intended score to calculate a baseline score for each question..... 71

Table 8: Relations Between Reported and Measured Learning Per LO..... 80

Table 9: Correlation between reported learning and Part Two of Post-Survey..... 83

Table 10: Correlation between measured learning and Part Two of Post-Survey..... 84

Table 11: All post-survey section 2 results and number of games played..... 110

Table 12: Complete data on participant changes towards the baseline. Measured per participant, per question, with aggregate scores. Participants who gave answers closer to the baseline after playing Obsolescence have positive scores depending on the degree 112

‘OBSOLESCENCE:’ EVALUATING AN EDUCATIONAL SERIOUS GAME ON ARTIFICIAL INTELLIGENCE IMPACTS TO MILITARY STRATEGIC GOALS

I. Introduction

Purpose and Problem Statement

Artificial Intelligence (AI) technologies have an almost unprecedented potential to change the shape of modern military conflict. However, the Department of Defense (DoD) lacks the educational capabilities required to prepare for changes caused by those technologies. There are neither enough resources to teach the concepts and values of an AI-saturated domain, nor are there adequate metrics to evaluate the potential educational use of Serious Games (SGs). Educational assessments of SGs often rely on unverified theories, subjective measurements, and anecdotal evidence.

The purpose of this research is to develop and test a new educational SG that can give DoD decision-makers appreciation and values for how AI technologies may interact with strategic warfare in the next 15 years. Supporting the game, this research creates and

utilizes a framework and methodology for more rigorously assessing the educational value of educational games.

Background and Motivation

Importance of AI Education

Advanced AI technology will cause extremely disruptive effects to many domains, including the military. The Department of Defense (DoD) and its international opponents have both acknowledged the potential AI can bring to the military landscape. The US Secretary of Defense (SecDef) has stated that AI, as a military technology, is "in a league of its own"; the Russian President holds the opinion that "whoever becomes the leader in [AI] will become the ruler of the world”; and the Chinese Community Party’s 5-year plan uses AI as a "leapfrog" technology to rapidly gain military superiority [1], [2].

The first step to maintaining AI superiority is education. Reports from the Executive Branch, the RAND Corporation, and the Joint Artificial Intelligence Center (JAIC) all conclude that the DoD requires more data, strategy, awareness, and education on the disruptive effects of AI technologies [3]–[5]. These strategic changes are necessary due to the rapid technological pace and significant disruptive potential of AI.

Studying Serious Educational Games

Educational science does not propose a single best medium to teach a particular subject. Current science says that the most effective teaching methods vary based on the individual being taught, the topic, and a myriad of other factors related to how the human

brain works [6]. Multimodal learning environments are defined as learning environments that use two or more different modes, or mediums, to represent the content knowledge [7]. Some educators use a combination of methods and modes for students to find the educational mode or medium that resonates best within their context. It also allows students to switch between information representations as their knowledge progresses. SGs are a modern, multimodal type of learning tool and can also be easily used to supplement a more traditional educational environment.

Further research is required for educators to confidently use SGs as teaching tools. A study by Defense Advanced Research Projects Agency (DARPA) in 2005 found that "the evidence of potential is striking, but the empirical evidence for effectiveness of games as learning environments is scant" [8]. Assessing the learning opportunity a game creates is a challenge, and the inherent depth and variability within games complicates generalizable results from experimental research. Adding further difficulty, SGs can teach skills or knowledge that are not easily measurable, such as communication, resource evaluation, or the language and framework required to fully utilize other material [9]. Therefore, the potential of educational games needs to be rigorously studied and explored.

Research Questions

This study asks two primary Research Questions (RQs):

- RQ1: Does the game Obsolescence teach its Learning Objectives (LOs)?
- RQ2: How does the measurement of learning compare to the reported learning?

Approach

The Game: Obsolescence

Obsolescence was designed and built to military decision-makers teach five LOs related to the military effects of future AI technologies. The game intends to teach lessons up to the *value* level of Bloom's affective taxonomy through simultaneous turn-based gameplay representing the global power struggle between military forces [10]. The game's LOs were chosen based on the consensus between several federally-funded studies on the future of AI in warfighting. The game's mechanics were based on several sources, most notably the SG *Hedgemony,* created by the Rand Corporation for military usage [11]. The game's educational value, in terms of its LOs, is derived from the adaptation of *Hedgenony's* mechanics and from the inclusion of game cards with AI-specific mechanics. To increase study participation, Obsolescence runs entirely in a browser using JavaScript and can be accessed by any device.

Study Methodology

Data collection involved a pre-post survey. The assessment of Obsolescence's educational value involved self-reporting from participants and a comparison of the pre-post survey data. The post-survey had three sections. Questions allowed answers on a five-point Likert scale ranging from *(1) Strongly Disagree* to *(5) Strongly Agree*.

The reported learning was measured purely through questions in post-survey Part One asking participants to rate their learning for each LO. Part Two asked them to rate their

engagement with the game, both regarding their enjoyment, ease of usage, and the time they chose to spend in-game. This section also allowed space for participants to give short answers to their game and study experience. Part Three duplicated the questions asked in the pre-survey. These questions were example scenarios of strategic decisions relevant to AI and the DoD. They were sourced from authoritative reports on the values the DoD should hold when dealing with AI technologies. The baseline scoring for these questions was calculated using a panel of Subject Matter Experts (SMEs).

The measured learning for each participant was evaluated based on the changes between their pre- and post-surveys. Learning was measured based on if participants answered the same question differently after playing Obsolescence, and if their new answer was closer to the established baseline. This measurement was used in conjunction with their self-reported learning to determine the overall educational impact of Obsolescence.

Experiment

The experiment recruited 48 participants from across the DoD. Any Federal employee was eligible to participate in the testing and evaluation of the game. Between 3 November 2021 and 12 January 2022, participants tested Obsolescence by accessing a weblink with credentials provided via email. The study asked all participants to begin by taking the pre-survey, then playing the game at least once, and finally concluding with the post-survey.

Scope and Limitations

Several factors limited the scope of this experiment. Firstly, Obsolescence was not designed for a particular schoolhouse or training purpose. This meant that both the game design and the experimental design were targeted towards a population that was inclusive of all ages, positions, and levels of knowledge. The scope of the material behind the game was therefore very broad, and the game explores higher-level concepts applicable to more than just a specific job or skillset.

This research was conducted entirely virtually. This influenced the hosting decisions and subject recruitment plan. The website hosting process had technical and procedural limitations, restricting Obsolescence to a single-player experience. In addition, the participants in this study are anonymous and their participation is voluntary. As a result, the research could not guarantee a minimum level of time or effort from all participants. The players interact with only the game and have no external motivation to learn the material, such as a grade, nor any community around the game or the material. The game and survey were designed with these constraints in mind, limiting the designed length of the game and surveys to accommodate casual interests or time commitments. A longer game and more in-depth survey may be able to generate more exact or detailed conclusions.

This research reached out to many organizations to obtain volunteers. A majority of those organizations were schoolhouses or institutes related to education or gaming. The subject population that chose to play Obsolescence may not represent the average

Federal employee, as the participants that were interested in volunteering were likely already interested in AI, education, or gaming.

The experiment was conducted over three months, and the entire research process took 18 months. The experiment and the study did not follow up with participants to investigate the long-term educational benefits of Obsolescence.

In addition, this experiment did not intend to create the most effective learning experience for the selected LOs. The way the experiment employs Obsolescence was designed entirely to get the most objective assessment of the game, not to create the optimal learning environment. Adding additional material, such as pre-reading, a video lecture, or a virtual instructor, would cast doubt on the actual source of the achieved learning and introduce additional independent variables. Therefore, as this work attempts to isolate the game as the only independent variable, it limited or excluded external materials.

Lastly, this experiment assesses Obsolescence against its designated LOs. It does not concern itself with any educational benefit Obsolescence has outside of the intended LOs. The game may teach other skills or have other educational benefits, such as time management, resource prioritization, learning theory, vocabulary related to AI, or general technology usage. Survey questions studying those effects were excluded as they were not the main focus of learning and to minimize the time commitment of participants.

Contributions

This thesis contributes to studies on gaming in education, SGs, and AI education. The research produces a novel educational tool, Obsolescence, that meets some of the recommended DoD requirements for training competency to the level required from an AI-saturated environment. It also provides infrastructure guidelines for other research with educational SGs. The procedures and methodology followed herein can be applied to create and host other web-based SGs within the DoD, both for other experiments and for employment in educational settings. Lastly, the work designed a standardized and scalable methodology to create survey questions and a baseline to assess an educational. This methodology may be applied to other games, both those designed as SGs and other games appropriated for educational purposes. The game and methodology further research towards the evaluation of LOs that are not easily measured or reliably self-reported, such as communication skills, situational awareness, or in this case, mindsets and values related to preparing for disruptive AI technologies.

Chapter Structure

Chapter 2 provides a review of the relevant literature in the fields of SGs, game design, AI, and educational game assessment.

Chapter 3 describes Obsolescence. This section contains a full description of the factors driving the game's LOs. It also discusses the rationale behind the selection of these particular LOs and the game mechanics related to each LO and both RQs.

Chapter 4 presents the methodology template used in the game evaluation and assessment. The methodology described in this chapter is designed to stand almost completely independent from the design of Obsolescence; it can be applied to the assessment of other SGs without significant reworking. This chapter also goes over the specifics of the infrastructure supporting Obsolescence and the experimental procedures.

Chapter 5 analyzes the experimental results. The chapter discusses the data obtained from the experiment and conducts an analysis comparing the reported value to the measured value of Obsolescence.

Chapter 6 concludes the paper with a summary of the work and a list of the research contributions. In addition, it puts forth areas of future work within both Obsolescence and the fields of SGs and educational game assessment.

II. Background

Chapter Overview

This chapter covers terminology and relevant literature related to utilizing SGs for education. SGs are defined as "any form of interactive computer-based game software for one or multiple players to be used on any platform and that has been developed to be more than entertainment" [12]. Current SG research has not conclusively proven any benefits of using a game over another type of educational medium, largely due to problems with assessing and evaluating learning from games [13]. However, SGs do have a place in education, especially as tools to help learners interact with the information at a deeper level. One area that SGs can be applied is in preparing the DoD for emerging AI technologies. This chapter discusses several sources that categorize the advent of the "AI Era" in military conflict and provide authoritative guidelines to manage such changes [14]. Lastly, this chapter discusses the tools used for this experiment and concludes with a summary of the background research.

Serious Games (SGs)

Definitions and Use Cases

SGs are usually implemented as aids to more conventional training or education [15]. This research is primarily concerned with Educational SGs. Educational SGs are designed with specific LOs. The game's designers, instructors, and mediators are aware

of the LOs and use the game as a medium through which they can transfer information. There is an important difference between educational SGs and wargames. Wargames, also categorized as SGs, explore potential futures via gameplay analogous to real-life, and thus allow players and researchers to make predictions on real events based on gameplay [16]. This research is not concerned with any wargaming aspects, only the educational benefits associated with LOs.

SGs create a learning environment where participant interaction is essential and that progresses with student engagement. Interaction and decision-making provide greater opportunities for players to internalize lessons, creating deeper and more effective learning [17]. Research into the educational benefits of SGs, especially digital SGs, is still in its infancy, but current studies show that SGs do not necessarily hold an overall learning advantage over other forms of learning [18]. However, these studies show that SGs "garner high engagement metrics, appeal to certain learning types, and work well for hand-picked modules" [18].

Research on SGs for education often draws from Flow Theory. Players in a game can experience a ‘flow’ state of complete involvement or engagement which has a positive effect on their learning [19]. This engagement also encourages longer training times and greater learning opportunities than other mediums [20]. Games also allow and require the immediate practice and application of the skills or lessons being taught. This not only engages players but enables the transfer of more complex skills and information [13].

Bloom’s taxonomy describes an understanding of educational mastery using layered structures to describe increased levels of learning [21]. Using Bloom’s framework, learning is divided into three categories, cognitive, affective, and psychomotor [10]. This research focuses on measuring education within the affective and cognitive domains, described in Figure 1.

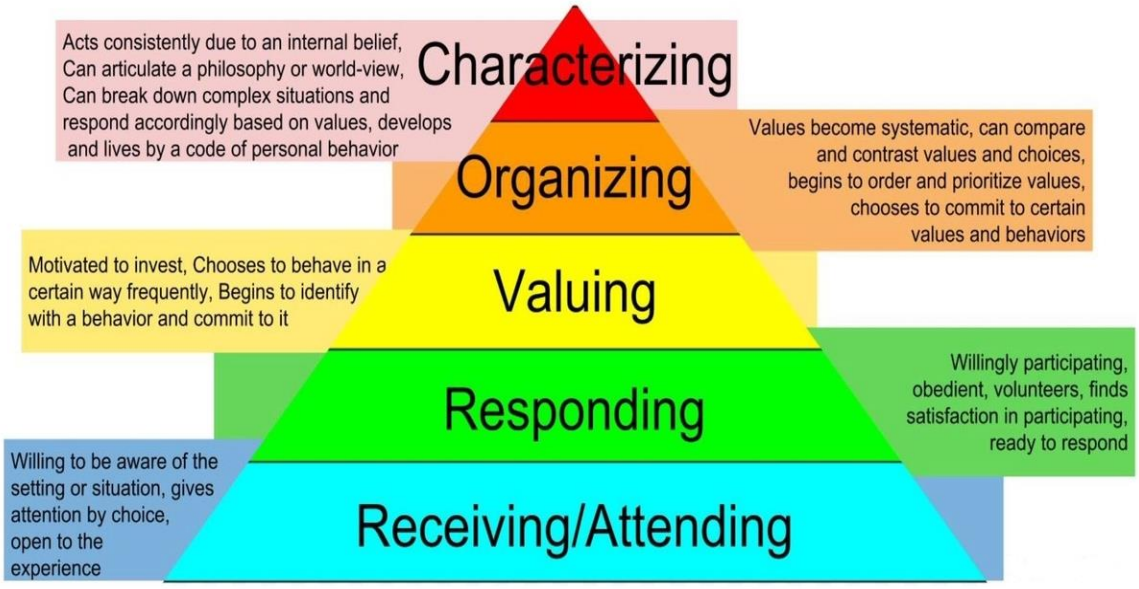


Figure 1: Bloom's Affective Taxonomy [10]

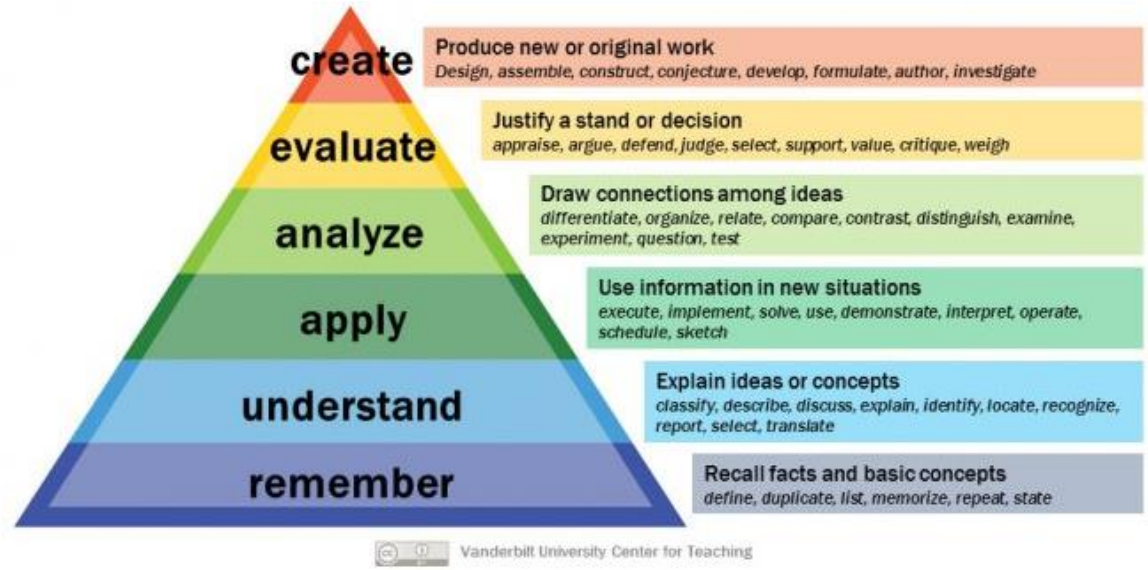


Figure 2: Bloom's Cognitive Taxonomy [21]

The affective domain involves feelings, emotions, and attitudes. It categorizes how information is internalized. First, a person must *receive* the idea, be aware of its existence and choose to pay attention to it. Then, they must *respond* in some way to the topic. Next, they should be able to see and express the *value* of the topic or idea. At a higher level, a person can *organize* different ideas and information to create their own value system. Lastly, learning is complete when a person can *characterize* their behavior by those values, affecting their everyday actions and becoming part of their self-definition.

The cognitive domain categorizes mental skills and knowledge. Like the affective domain, higher levels build off of the abilities from lower levels. The scale measures abilities from basic recall to the production of novel work.

Low levels of affective learning can be relatively simple to assess by testing awareness of a concept or the ability to logically explain aspects of its value. SGs may effectively teach to the *valuing* level or higher, as they can create situations where the player has to commit to valuing certain traits, concepts, or information to succeed [13]. As an example, medical SGs have been used to provide training on Clinical Reasoning, a skill set encompassing proper diagnosing, institution of appropriate treatment, and managing emerging complications [22]. The SG provides the practice and hands-on experience that encourages values positive to clinical settings and requires players to act on their own beliefs and values. Successful players will have to, in-game, commit to and live by certain values and behaviors. An effective educational SG would transfer in-game learning to real-world application, influencing how participants value, organize, or characterize complex or abstract topics.

Likewise, cognitive learning can be easily assessed at a low taxonomy level, and difficult at a high level. Tests graded based on correct answers can measure knowledge of facts, identification of terms, and some application of terms. For instance, a math test may measure a student's ability to *apply* information to a new problem. However, measuring a student's ability to *analyze, evaluate,* or *create* would likely require an instructor to determine, subjectively, if the student was demonstrating those abilities.

SGs in Education

SGs are usually used in conjunction with other educational methods and are rarely the sole source of information students receive. Some sources show "the real potential of educational games is realized only when teachers join students in interacting" [15]. In a

2015 literature review on SG evaluation, most of the applications of educational SGs applied the SG within a classroom or alongside a similar educational setting [23].

SGs can also provide other benefits. The DoD has shown interest in implementing SGs both to make use of their multimodal nature and to improve the course development and deployment timelines [24]. Most educational courses have long development and implementation timelines, which increases if the course in question is digital [25]. According to an Acquisition Education Research Analyst for the Air Force Institute of Technology (AFIT), a standard "informal education product" takes "a period of a few months to a year" to develop [26]. As of 2020, a digital SG similar to Obsolescence would take an estimated 155 hours, or about 4 weeks, for a professional team to develop [27]. That time does not include the course development work surrounding the game; however, even if the game does not reduce any of the normal course development work, the addition of 4 weeks of work (for one person) would not significantly alter current timelines.

Current State of SG Evaluation and Assessment

Assessing the educational value of a specific tool or methodology is difficult with any medium, but SGs, in particular, have several additional challenges. Implementing SGs within a course or alongside other types of learning is very common but makes rigorous assessment and evaluation of the game more difficult. Multimodal learning is an effective educational strategy [6], [7]. However, when conducting

assessments of one piece of the process, each dimension of the environment can become an unwanted independent variable. To effectively evaluate an SG, and only the SG, the game must be able to function as a stand-alone educational tool.

The intended purpose of an educational SG includes teaching the material in a fun, entertaining, or engaging fashion. Assessments of SGs present unique challenges as the goals of education and enjoyment are often entangled. Research has shown that high engagement levels strongly correlate to the amount of reported learning [28]. Studies have also shown that engaging SGs hold subjects' attention for longer and create higher levels of intrinsic motivation [28]. A full SG assessment, therefore, covers an engagement assessment of the players and an assessment of the LOs [29].

The vast majority of SG studies conduct educational measurements via a pre and/or post-test developed specifically for the game [18], [23]. Figure 3 outlines the prevalence of questionnaires for measuring the educational value of SGs. Other measurement techniques include interviews, game logs, discussions, and observations of the game session. Survey questionnaires have been used to assess both game enjoyment and educational value, and game logs can provide direct measurement to support the assessment results.

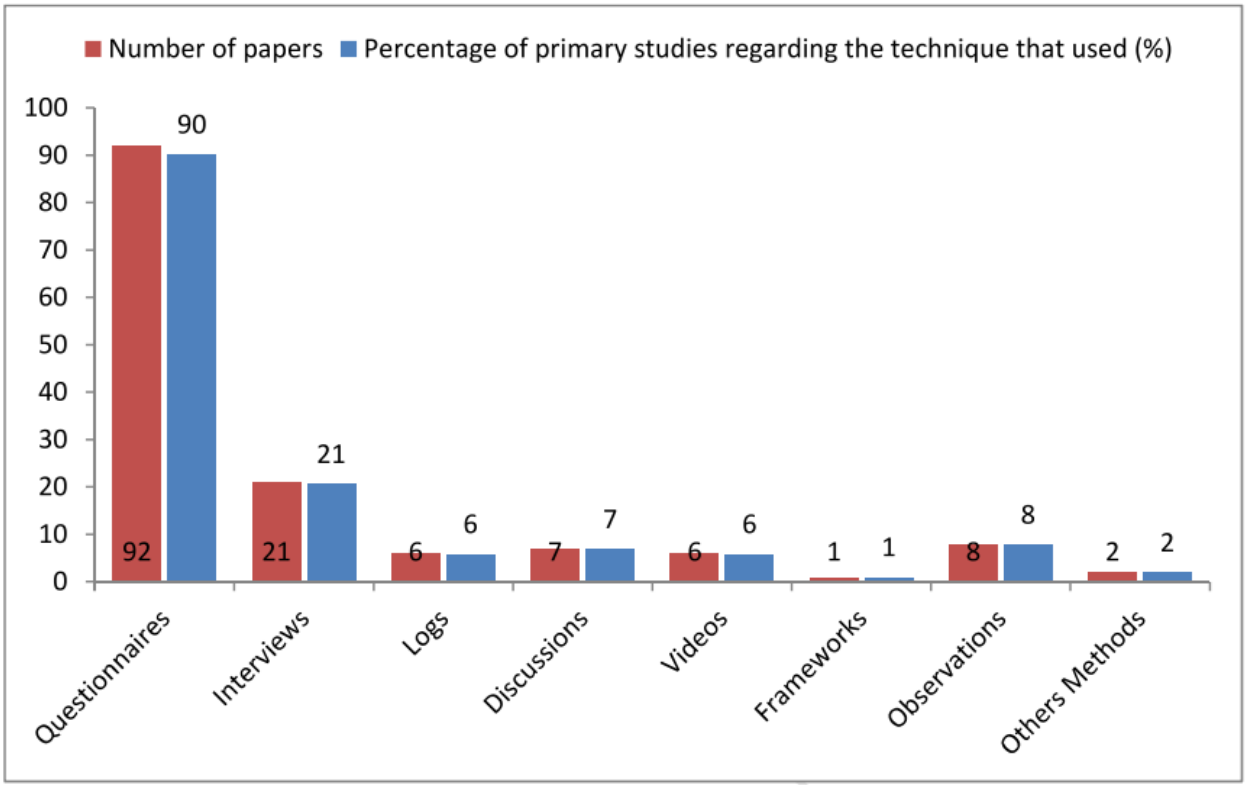


Figure 3: Techniques used to evaluate SGs, in number and % of 102 papers [22]

When evaluating surveys, most research (85.2%) uses Likert scale questionnaires [18]. Of the studies that utilized questionnaires, the majority use only a post-test, either measuring reported levels of learning and engagement, or lower levels of learning such as memorization or definitions. Figure 4 shows a measurement of the most types of game assessment surveys [23]. Some studies utilized multiple post-tests for longitudinal research, and only 15 out of the 102 used some form of a control group or baseline to evaluate their answers [23].

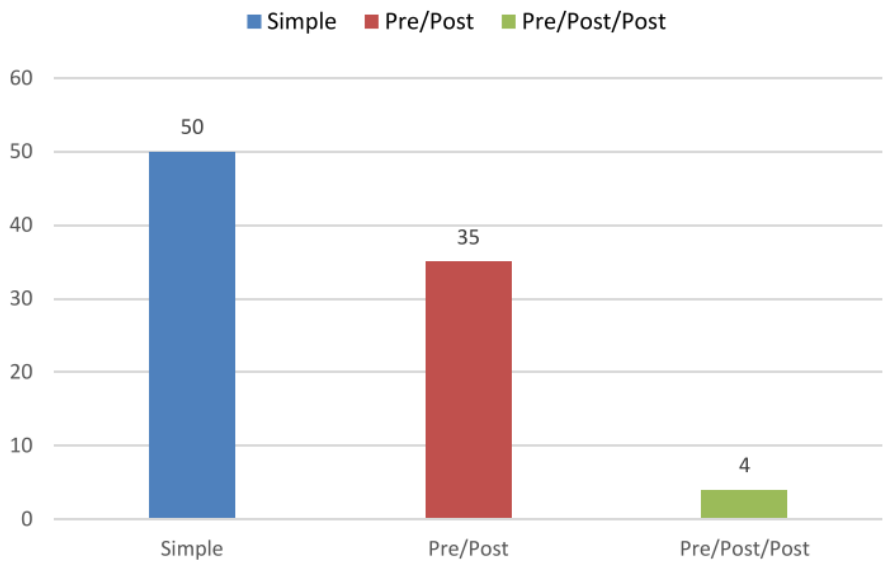


Figure 4: Common Questionnaire types [22].

When a pre and post-test are both implemented, the educational success of the SG is measured by the change in performance. In 2013, a study conducted on those types of evaluations found no generalizable and consistent result across all SGs; the researchers were not able to conclude anything about SGs as a whole [28]. Many of the games studied either had no significant learning effect or were comparative to a control group using a different medium [28]. This supports other findings indicating that while SGs may be an effective tool, research has yet to find consistent results that can generalize to all SGs, or settle on a particular evaluation methodology [13].

Most SG studies have small participant populations. Figure 5 shows the population sizes of 102 SG studies with two scales, 1-120+ and 1-40. The majority of studies test the SG using less than 40 participants, and the most common size is between 11 and 20 people [23].

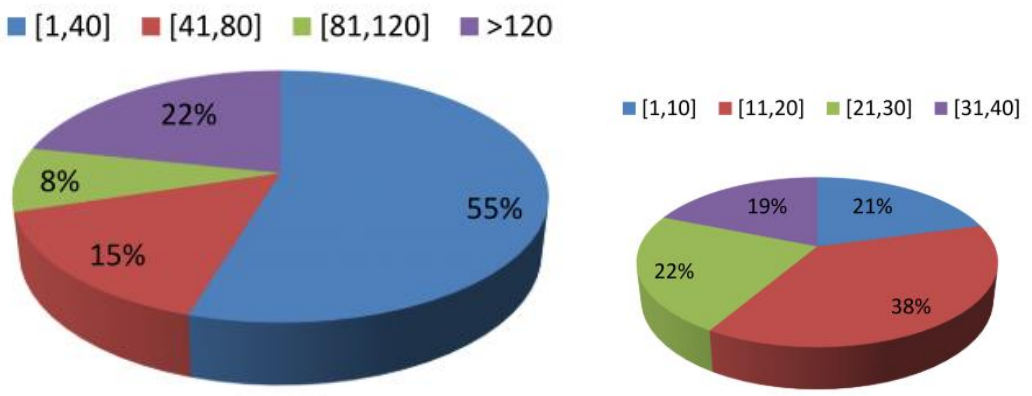


Figure 5: Distribution of SG study population sizes, displayed both between 1->120 participants and 1-40 participants [22]

Studies on SGs have not determined the exact educational differences between SGs and other mediums. A 2005 review of SG literature found that only 19 out of over 4,000 published, peer-reviewed articles contained either qualitative or quantitative data from an assessment of learning or motivation [8]. However, those studies that did conduct a scientific evaluation were usually found to have significant threats to their validity [8]. In 2007, Richard Clark cataloged the following major problems almost all positive results from SG research "tends to ignore" [13]:

1. *Evaluating only post-game knowledge:* without a pre-test of some sort, participants may just be demonstrating their prior knowledge and abilities.
2. *Evaluating games without a scientific control or baseline:* many evaluations compare the learning from a game to a control group that engaged in an unrelated activity or had no instruction.
3. *Confusing Educational SGs with Wargames and simulations:* the terminologies and definitions surrounding SGs are also used in the study of related constructs, leading to occasional confusion when interpreting studies.
4. *Evaluating games based solely on reported opinions on learning and motivation:* a majority of studies do not implement direct measures of

learning, which often conflict with self-assessments when both are gathered.

5. *Designing SGs without grounding in pedagogical methods:* educational games that employ self-guided, discovery, constructivist, or problem-based learning pedagogy are less effective than games designed with direct instructional methods.

SGs may not be the best choice for educators to use in every situation, but even without exact methods of assessment, researchers believe SGs have a place in education. In a literature review of SG design and evaluation, De Gloria et al. identified several challenges mitigating the effectiveness of SGs [30]. SGs typically require a "suspension of belief" to get immersed in the game and the game's mechanics [30]. SGs, especially digital ones, can cause frustration from usability issues. This is exacerbated by the term 'game'; many commercial games cost immense resources to create and polish, so potential players might begin a SG expecting a similar level of investment. Competitive aspects, while sometimes motivating, can cause frustrations in some players and detract from the educational value. Despite the listed challenges, the survey concludes that SGs are effective and have huge potential, especially as the tools for designing, constructing, and evaluating games continue to grow [30].

Disruptive Artificial Intelligence Technologies

Obsolescence, a digital SG, teaches militarily relevant mindsets and values for interacting with disruptive AI technologies in the next 15 years. The 2020 report from the National Security Commission on AI (NSCAI) heavily stresses AI-readiness [14]. They put it quite starkly: "Our armed forces' competitive military-technical advantage could be lost within the next decade" if the DoD does not "achieve a state of military AI readiness

by 2025" [14]. One of the first steps in this process is to ensure the Joint AI Center (JAIC) builds a roadmap towards AI integration for the next 5 years [4]. The JAIC acknowledges the poor state of the military in terms of AI posture and believes that the "DoD must prioritize education and training… to deliver AI capabilities" [5]. Their planned training covers both basic AI literacy and also includes strategic-level competency. As mentioned above, SGs have the potential to teach to high levels of comprehension without taking undue development time. SGs focused on AI might fit perfectly into the educational plans of many DoD organizations.

Key sources in the development of Obsolescence and the creation of the game's LOs are summarized here:

U.S. Military Investments in Autonomy and AI: A Strategic Assessment [31]

Created by the Center for Security and Emerging Technology (CSET) for the DoD, this assessment examines the scope and implications of U.S. military investments in autonomy and AI. It focused on AI technologies, critical capabilities enabled by AI technologies, and the strategic ramifications from judicious and non-judicious applications of those capabilities. The report contains short and long-term recommendations for different parts of the DoD. Their first recommended action is to "fill knowledge gaps" about what AI will and can mean for militaries [31].

Preparing For The Future Of Artificial Intelligence [3]

Created by the National Science and Technology Council (NSTC) Committee on Technology for the Executive branch, this report is a survey of the current and potential

state of AI applications and the impacts on society and public policy caused by AI technological advancement. The document covers all sectors of the government and specifically explores military concerns, primarily in the fields of cyber security and autonomous weapons systems. The document includes a list of recommendations for changes for high levels of the Federal government and the DoD.

DoD AI Education Strategy [5]

In this document, the JAIC outlines its first steps towards making the DoD an AI-capable force. The DoD is competing globally for AI talent and is "not yet postured to compete with industry in hiring" [5]. To solve this, the JAIC prioritizes education across the DoD to create AI talent from within the workforce, and to have members of the DoD mesh seamlessly with contracted AI experts. The strategy is broad but includes specific measurements of success that certain populations of the DoD should meet by the end of their respective training pipelines.

The Department of Defense's Posture for Artificial Intelligence: Assessment and Recommendations for Improvement [4]

Created by the RAND Corporation after a request by Congress and the JAIC, this document studies what changes the DoD needs to make to take advantage of emerging AI technologies and avoid safety risks. It addresses DoD decision-makers at a strategic level and does not assume any prior knowledge about AI. The research first analyzes the DoD's current posture for AI, then provides a series of 11 recommendations.

mmowgli - Design for Maritime Singularity: Final Report [32]

This research supports the Office of Naval Research, Director of Disruptive Technologies. The study used mmowgli, an online platform used for conducting large-scale research, to explore how the U.S. Navy might respond to a future scenario often described as the Singularity. They posit two scenarios, each a different definition of the Singularity, and asked the mmowgli population to contribute to brainstorming and forecasting probabilities. Following the online session, an in-person workshop refined the ideas into actionable recommendations. Their recommendations are focused on helping the Navy address likely and worst-case scenarios related to disruptive AI technologies and can be generalized to all the U.S. military branches.

- **Final Report: National Security Commission on Artificial Intelligence** [14]

In 2020, the National Security Commission on Artificial Intelligence (NSCAI) began a report on "the development of AI, machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States" [14]. The comprehensive document they produced primarily discusses international and military implications from either advances in AI technology or more widespread adoption of currently existing technology. The report is broken into two sections, the first discussing "Defending America in the AI Era" and the second "Winning the Technology Competition" [14]. Their report describes the behaviors and mindsets that are required by a military competing with and against AI capabilities.

Tools

The following tools were used in the development and experimentation for this paper:

GameMaker Studio 2 – HTML5

GameMaker Studio 2 is a game development environment specialized to enable producing two-dimensional games quickly [33]. Code is written in GameMaker Language (GML), which is syntactically very similar to python but is entirely object-oriented. The environment supports exporting from GML into JavaScript and HTML5, as well as locally hosting web servers for testing purposes. This research used a personal Gamemaker license to enable exporting to HTML5. GameMaker was chosen as it fully supports 2D games like Obsolescence and has a short learning curve.

Microsoft Azure, Docker, and Apache

The experiment was hosted using a combination of cloud services from Microsoft Azure and a Docker container with an Apache webserver. The cloud service allows for remote database management and for automation of infrastructure tasks, such as compliance checking, automatic storage and compute scaling, and development pipelines. It allowed the source code for Obsolescence to be uploaded and modified through an automatic system accessible from any internet-connected device. The HTML5 game files were served by an apache server running a simple authentication protocol. This server was virtualized and contained using docker, allowing modularity, duplication, and the

ability to save run-states. The container saved cost and allowed for a standerdized format within the cloud space regardless of what application is running.

Background Summary

SGs are used as an educational medium that increases participant engagement, time spent learning, and hands-on experience with the information. Due to those effects, learning theories posit that SGs can be a more effective teaching tool when applied correctly. However, neither positive nor negative effects of SGs as a medium to promote learning have been confirmed. Evaluations of SGs suffer from variability between games, population sizes, an excess of confounding variables, and inexact measurement methodologies. Despite those issues, SGs are a promising tool for an alternate mode of learning and can be powerful when used properly and in conjunction with other educational methods. The ramifications of disruptive AI technologies are one such place that educational SGs may be useful, given the recommendations of more abstract and higher-level learning. AI technologies are likely to create significant disruptive effects in military functions, and a variety of authoritative sources agree that the DoD should start addressing necessary changes with education.

III. Design of Environment

Chapter Overview

This chapter describes the LOs and associated design decisions of the game Obsolescence. It covers aspects of the game from the initial motivation to the technical details of the online game's hosting.

Obsolescence is a digital board game whereby players play as competing militaries attempting to gain the most Influence Points (IP) over a set number of turns, representing years. Players develop and move military forces, represented by tokens, around a map of the globe. Instead of military forces directly fighting each other, conflict is represented in the form of dominance struggles. Every turn, the player with the most forces in a given region gains IP, representing that they can achieve whatever their military/political goals are and hinder their adversaries' goals. Players plan their moves simultaneously, and when all planning has been completed, all force movements happen simultaneously and the scores for the round are added to each player's total.

Targeted learning and variation between individual games comes from Technology Cards (Tech Cards). Tech Cards represent an AI-based technology that militaries can choose to adopt. As the game progresses, more technologies will become available to all players. When adopted, each Tech Card gives unique abilities or benefits. Players, therefore, compete by building and moving forces around the map while allocating resources to adopt a set of technologies that gives their forces a critical edge. The rest of this Chapter discusses Obsolescence's LOs and details the gameplay and

infrastructure design choices. A more complete description of the game mechanics and rules can be found in Appendix A.

Obsolescence's Learning Objectives (LOs)

Rationale Behind LO Selection

The DoD employs a wide array of think tanks and runs several organizations dedicated to strategic policy guidance. For Obsolescence, LOs were derived from publications based on the authority, completeness, and relevance to the DoD. The Federal Government also commissions frequent reports about issues that overlap with military interests. The publications chosen all had specific recommendations or laid out objectives for the DoD related to the future of AI technologies. These were analyzed and clustered into 8 general recommendations for the DoD as an organization. From those, they were further refined into the 5 LOs based on the feasibility of implementation into a game, and the perceived weight given to them from the report. The three potential LOs that were not selected are as follows: *recognize that data is a key resource for successful military ops in a world with advanced AI*, *recognize that the supply of 'compute' is critical to advanced AI*, and *identify potential DAI technologies that require monitoring*.Table 1Table 1: List of Learning Objectives (LOs) for Obsolescence, alongside their intended Cognitive (C) and Affective (A) Taxonomy level and source documents.

Obsolescence was designed to be able to teach players each LO up to a certain taxonomy level. It focuses primarily on the affective domain, influencing players' opinion of the worth of several AI-related concepts. All LOs except for LO2 fall within

taxonomy level 2, *valuing*. The game also attempts to teach cognitive concepts to the *evaluate* level. Players learn an understanding of the game, apply in-game knowledge and concepts as they progress, and must weigh choices based on their own judgment of worth. However, while Obsolescence and most SGs can teach in-game concepts to a high taxonomy level, this does not guarantee that the LOs, which are real-world concepts, are taught to the same level. Players need to make logical connections between in-game and real-world values and decisions.

Table 1 contains the list of intended LOs for Obsolescence. It also contains the maximum intended Taxonomy level to which the game is designed to teach the LOs and the sources from which the LOs were derived. Verification, Validation, Testing, and Evaluation (VVT&E), referred to in LO1, includes all activities intended to ensure a particular technology performs as intended and without safety concerns [4].

LO#	Task	Taxonomy Level	Sources
1	Recognize and defend the value of VVT&E for all disruptive AI technologies	(A) 3 (valuing) (C) 5 (evaluate)	[3], [14], [31], [4]
2	Recognize that disruptive AI technologies would greatly increase the complexity of the military environment	(A) 1 (receiving) (C) 2 (understand)	[14], [32]
3	Support and value increases to military 'Complexity Carrying Capacity'	(A) 3 (valuing) (C) 5 (evaluate)	[3], [14], [32]
4	Assess value of strategic plans and roadmaps that deal with disruptive AI	(A) 3 (valuing) (C) 5 (evaluate)	[4], [5], [14]
5	Support and value increases to international monitoring and restrictions on AI progress and development	(A) 3 (valuing) (C) 5 (evaluate)	[4], [14]

Table 1: List of Learning Objectives (LOs) for Obsolescence, alongside their intended Cognitive (C) and Affective (A) Taxonomy level and source documents.

Obsolescence was designed to be able to teach players each LO up to a certain taxonomy level. It focuses primarily on the affective domain, influencing players' opinion of the worth of several AI-related concepts. All LOs except for LO2 fall within taxonomy level 2, *valuing*. The game also attempts to teach cognitive concepts to the *evaluate* level. Players learn an understanding of the game, apply in-game knowledge and concepts as they progress, and must weigh choices based on their own judgment of worth. However, while Obsolescence and most SGs can teach in-game concepts to a high taxonomy level, this does not guarantee that the LOs, which are real-world concepts, are taught to the same level. Players need to make logical connections between in-game and real-world values and decisions.

LO1. Recognize and defend the value of VVT&E for all disruptive AI technologies

After playing Obsolescence, participants should demonstrate abilities up to the *value* and *evaluate* levels [10], [21]. When presented with example scenarios participants should display increased value and prioritization for Validation, Verification, Testing, and Evaluation (VVT&E) efforts. Participants should select choices they or others make to invest resources into VVT&E for potentially disruptive AI technologies.

VVT&E for AI technologies is a common topic in the literature surrounding military usage of AI. Multiple sources stress the point: AI systems need VVT&E to be effective and low-risk [3], [14], [31]. Ensuring the military employs robust and effective VVT&E processes mitigates some of the largest roadblocks with new technologies such as wasted effort and cost, novel ethical concerns, and correct application in the field.

VVT&E efforts are even more important for AI technologies than for traditional military systems.

AI technologies have unique reasons for VVT&E, compared to other emerging technologies. First, many advanced AI technologies fall in the uncanny valley of comfort. Human operators naturally anthropomorphize AI systems, causing either over-reliance or over-confidence in the system, or false assumptions about how it works. AI algorithms can be often explained using simple human terms; however, this creates problems for engineers attempting to conduct comprehensive VVT&E processes, as these summaries might hide important differences.

For instance, AI systems are often described as having a goal [34]. An automatic sorting system might have the goal of sorting various balls into correct bins. However, at its core, the system is optimizing a set of parameters to minimize the number of reported errors. While this works well in practice, if a single bin's error detector fails, the system will rapidly learn to put every ball in that bin, as errors are never reported. The goal-based understanding hides emergent behavior that humans would not inherently expect. This can make evaluating AI systems more difficult if the evaluation framework does not demand rigorous procedures created by experts who understand how AI works.

In addition, neural networks are the foundational technology supporting many proposed AI capabilities. One of the significant disadvantages of such systems is that most neural net code and decisions end up unreadable to humans. This means that even the developers do not know exactly what formula the AI is using to make its decisions; it is almost impossible to guarantee performance or safety in a novel situation.

AI, in the modern era, is rapidly demonstrating proficiency when used as an expert system for a wide variety of domains. AI has beaten the world champion in Go, professional E-Sports teams, and military applications such as dog-fighting and aircraft detection systems [35], [36]. However, overreliance on AI interpretation, presentation, or judgment is already an issue for today's force [37]. Unless the systems are perfect, they must not be treated as infallible, regardless of how much better they can perform. Generals that rely on an AI-generated map of forces to make battlefield decisions need to understand the margin of error between the AI and real-life [37].

AI also requires stringent VVT&E efforts because of its role as a force multiplier. AI rarely stands on its own, but instead augments existing systems or processes. This can transform small functional or ethical issues into significant errors or scandals. For instance, in 2019, lawyers discovered that an AI algorithm deployed in US hospitals with over 200 million patients did not train on data completely cleared of all racial indicators. As such, the AI was heavily favoring white patients over black patients for extra medical care [38]. This problem resulted from insufficient VVT&E, likely stemming from a neural net that was rapidly deployed only after assuring that it met the bare minimum requirements.

Lastly, a significant group of AI researchers predicts that within this century advanced AI will be an existential risk to humanity exceeding global nuclear war [32], [39]–[41]. Even a small percentage chance of a disaster of that magnitude warrants extremely careful consideration when developing and employing such technologies.

LO2. Recognize that disruptive AI technologies would greatly increase the complexity of the military environment

After playing Obsolescence, participants should demonstrate abilities up to the *understand* and *receive* levels of Bloom's taxonomy [10], [21]. Participants should be aware of the effects that disruptive technology can have on the information environment for strategic military decision-making. When given sample scenarios, participants should recognize the potential complexity of disruptive AI technologies and support efforts to increase awareness of the effects. An understanding of this LO is critical for the higher taxonomies of learning taught by LO3.

It is increasingly difficult to understand a single military situation completely. The world is becoming more interconnected, with technology and society building off of earlier foundations. Shops in rural America now compete with big businesses in east Asia, and military decisions made in Western Europe have potential ramifications in South America. Furthermore, governments and individuals can now capture and ingest increasingly larger data sets. Commanders can see live video streams of troops in combat and can talk in real-time with their peers across the globe to seek optimal strategic decisions. This influx of information and options does not always help decision-makers but can create situations of extreme micromanaging or tunnel vision on a specific tactical objective [37].

LO3. Support and value increases to military 'Complexity Carrying Capacity'

After playing Obsolescence, participants should demonstrate abilities up to the *value* and *evaluate* levels of Bloom's taxonomy [10], [21]. Participants should, after understanding how the military environment is rapidly becoming more complex, LO2, value capabilities and solutions that give decision-makers abilities to deal with large or complex information sets.

While AI is increasing the complexity of the military environment, it is also providing solutions to compensate and enable modern warfighters to operate even more efficiently. AI can bring significant increases to the complexity carrying capacity by distilling, displaying, and analyzing, data now being collected at such a large scale that human operators cannot keep up. This is likely a more disruptive effect from AI technologies than robotic vehicles or autonomous weapons. Technologies that improve what humans can already do are generally not as disruptive as technologies that bring novel capabilities. AI systems are specifically optimized to operate within vast amounts of data.

Multiple sources warn that this could reach a point where decision-makers do not have the time or ‘complexity carrying capacity' to effectively make decisions [14], [32]. AI technologies, in particular, are characterized by some as the next industrial revolution [42]. The massive amount of data and power enabled by the internet may only be fully realized with scalable intelligences designed to work within that framework. Military decision-makers need to recognize the changing terrain and adapt their mindsets, priorities, and strategies accordingly.

AI can be applied to analytics, making decisions, or carrying out a task. But perhaps more importantly, it can be used to automatically distill and display relevant information. A military AI advisor could be aware of every single event the DoD was tracking, and selectively display relevant summaries of pertinent events to any topic a commander queries. This, according to research from NPS, might prove to be one of the biggest strategic advantages militaries can expect from AI in the near future [32].

LO4. Assess value of strategic plans and roadmaps that deal with disruptive AI

After playing Obsolescence, participants should demonstrate abilities up to the *value* and *evaluate* levels of Bloom's taxonomy [10], [21]. When given sample scenarios, participants should demonstrate stronger weights and values for proactive measures dealing with potentially disruptive AI technologies at a strategic level.

This is a skill set that the DoD needs more of, and not just for AI technologies, but all of Information Technology (IT). The first Chief Software Officer for the USAF, Nicholas Chaillan, said the following concerning the DoD's current management of IT and software projects:

> We would not put a pilot in the cockpit without extensive flight training; why would we expect someone with no IT experience to be close to successful? They do not know what to execute on or what to prioritize which leads to endless risk reduction efforts and diluted focus [43]

This opinion is just as applicable to AI as it is to IT. Decisionmakers, and especially future decision-makers, need to adopt a new perspective considering the future of AI technology. Any strategic plan that projects over 10 years into the future needs to

include preparations for likely technology changes. These plans should also consider unlikely, but highly damaging, technology developments. AI research and development is hard to predict with accuracy. In the interests of national security, DoD decision-makers need to prepare to critically analyze and evaluate predictions and roadmaps involving the future of AI technology.

LO5. Support and value increases to international monitoring and restrictions on AI progress and development

After playing Obsolescence, participants should demonstrate abilities up to the *value* and *evaluate* levels of Bloom's taxonomy [10], [21]. When evaluating sample scenarios, participants should show increased support and value for the international monitoring and restrictions of AI technologies.

The American military needs to prepare for wars fought with future technology. Many experts agree that future wars will be shaped by advanced AI technologies. It is critical, therefore, to invest significant resources into both monitoring and regulating international AI technologies, especially as related to warfare. AI technologies can easily cause disproportionate ethical harm. One military goal is to avert potential international crises before they even occur. As a potential cause of many such crises, and in addition to their own military ramifications, AI technologies should be a military intelligence priority.

Design of Game Mechanics

Obsolescence was designed to represent a simple model drawn from real concerns of what the highest-level decision-makers in the DoD might do/see. The game is meant to start at the current year and progress up to 15 years in the future. As players play the game, they should realize how seemingly low-level AI technologies can drastically change even the highest level of military objectives. The overall structure of the game was influenced by the game *Hedgemony* produced by the RAND Corporation [11]. The specifics of the design were, in large part, focused on a US-centric view. Table 2 describes the similarities between Obsolescence and Hedgemony.

Game Design	Obsolescence	Hedgemony
Resource types	Resource Points (RP), adopted Tech Cards	RP, Force Mod level, Critical Capability Mod level, National Tech Level
Victory condition	Have the most IP	Have the most IP
Force Abstraction	Tokens represent strategic level capabilities	Tokens represent strategic level capabilities
Scale and Scope	Play as opposing nations' militaries	Play as opposing nations' militaries
Available actions	Move forces, build forces, interact with tech cards	Move forces, build forces, develop forces, conduct diplomatic actions, interact with action cards, other actions per Game Mastes' discretion
Game change over time	Tech cards are adopted, changing game rules	R&D level increases, Game Master scenarios progress
Modularity of game	Game settings can be adjusted	Game Master can set up specific scenarios for games, or during games
Opponents	All AI opponents	Other players + Game Master(s)
Multiplayer	Singleplayer only	Multiplayer only
Game completion requirements	Set number of turns	Set number of turns
Teams	Singleplayer only	Competing teams of supporting nations

Feedback	Post-game scores	Group Discussion among players and Game Master breaking down game events
Tabletop or virtual	Virtual	Tabletop
Asymettry	All players start equal but get different random objectives	Nations start with different capabilities and objectives
Time pressure	In-game turn timer	No time pressure
Game Events	Random Objectives, random available Tech Cards	Shuffled decks of potential actions, scenario-specific events

Table 2: Comparison of Obsolescence and Hedgemony[11]

To build off of previous work creating a realistic, strategic level military game, most elements of Obsolescence were designed either to replicate the corresponding element of *Hedgemony* or to simplify its game design. The most significant exceptions are the exclusion of a game master and the choice to make Obsolescence singleplayer. These decisions were motivated primarily by two reasons. As a video game, Obsolescence is not able to implement a Game Master or collaborative team play as easily as the tabletop game *Hedgemony*. Secondly, Obsolescence was built with this research in mind, and therefore game elements were optimized for clarity of analysis. This motivated the removal of potential confounding variables such as unstandardized Game Master behavior and multiplayer dynamics.

Figure 6 displays a screenshot of a game in progress, with labeled interface items. The game is on turn 10, following the China player. They have developed forces and deployed them globally to several regions. They have also adopted four AI technologies from the cards available at this point in the game, granting them additional passive and active abilities, such as the ability to see projected enemy movements. With their remaining 3 Resource Points (RP), the player can move forces to achieve local

dominance in a region, develop more forces, including their new *drone swarms* technology, and/or take actions to adopt new Tech Cards.

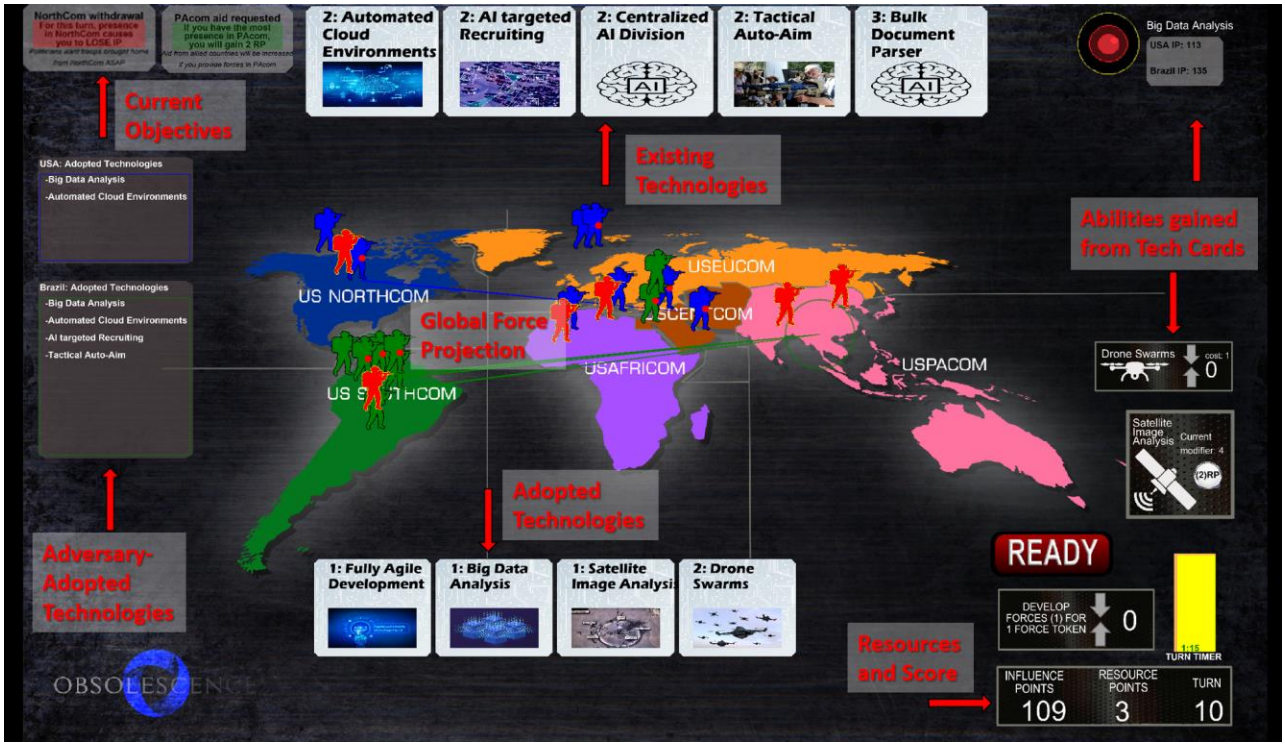


Figure 6: Sample Obsolescence screenshot with labeled interface icons.

Obsolescence strives for realism in the following ways. Geographical regions are not conquered, multiple opposing nations can have a military presence in the same geographical area. Similarly, players do not strive to destroy their opponents' units but to render them strategically ineffective. Instead, Influence Points (IP), the game's win-condition, are used to represent the vague quantifier of how well any military achieves its highest-level objectives. These represent how well the military achieves the political goals of its country, serves its people, and is prepared to defend its nation.

Despite the game simulating players as the highest level of military command, players do not choose their objectives. This influence from outside factors represents real-world social, political, and cultural factors that put an impetus on militaries to conduct certain operations or refrain from taking certain actions.

Multiple aspects of the game are significantly abstracted from their real-world counterparts. Force Tokens are purposefully abstracted out from a specific military unit. The definition of military forces has become vague in the 21st century when wars can be waged by non-uniformed personnel or as massive false-flag operations. The abstraction also allows for units such as cyber forces to have an in-game representation. A single Force Token can represent any combination of military assets. In part due to this abstraction, and partly due to the DoD's global logistical system, geographical adjacency is not a determent factor in Force Token movement. Resource Points are used to represent budget, policy priorities, manpower, and any other limited strategic level resource. The geographic map only displays US Combatant commands. The simplicity in this level of abstraction prevents overly-complicated gameplay.

The Tech Cards in the game represent specific AI-related Technologies with military relevance and imitate realistic technology adoption through two steps. First, the technology needs to reach the point in development to be usable. Second, the technology needs to have a military invest in the technology and begin using it in operations.

In AI industrial base, the majority of cutting-edge development is not for military-specific usage. Therefore, to model the current state of military usage of AI technologies, the players have no control over which technologies are developed enough to be used.

This adds a bit of timeline uncertainty, a common theme in any predictive AI research. The uncertainty and randomness force players to either play reactively or prepare for a possible emergence of any of a dozen technologies. Table 3 displays all the Tech Cards utilized in Obsolescence. Each Tech Card has an associated LO, a timeframe where it can appear in the game, a resource cost, and its effect. Each Tech Card also has additional flavor text describing a theoretical military usage of the AI technology.

Tech Card Name	LO(s)	Time-line	Cost	Effect	Flavor Text
AI Testing & Evaluation	3,1	2	3	Automatically VVTEs all techs for free.	Software already can conduct many 'quality assurance' and security audit functions. AI software will likely give a better estimate on a novel system's reliability, security, and projected affects than humans can, especially as novel systems get more integrated and complicated.
Robo Logistics	2,3	2	5	Each force movement costs 1 less	Robotic cars, boats, factories, delivery systems, and (perhaps most importantly) inventory tracking systems: correctly implemented automated logistics systems can save incredible amounts of time, money, and manpower, especially for a multi-trillion dollar organization.
Strategic assistant	3	2	3	Gives you infinite time to take your turn.	As the information era progresses, higher-level leadership will get more and more inundated with 'critical' information. A strategic level AI to augment decision-making can clarify situations and data sets, allowing swifter and more assured decisions.
AI Induced Radicalization	3	1	4	Every turn, for each green and blue objective, develops a free force token already in the target COCOM. At the start of every turn, retires an additional random force.	Advanced chatbots can be given agendas to incite local riots and militias- essentially acting as your own military force in another territory.
Centralized AI Division	4	1	4	All active AI effects cost 1 less. VVTE	Having a centralized (likely cloud-based) location for AI technologies, and tying it

				actions cost 0.	into a similarly centralized military organization greatly allows the elimination of redundancies and extreme cost savings, in addition to the strategic benefits centralized command always had.
Drone Swarms	2	1	5	Every turn, lose 1 IP for each Drone Token you control. Spend (1)RP: build a 'Drone Token' that functions as a Force Token but with 3x the strength.	Once the AI for drone swarm control has been built, fleets of weaponized and tiny quad-copters are arguably the most cost-effective way to project force
Bulk Document Parser	5,3	1	5	Can see other players RP	AI software is getting better and better at understanding the written word- and what it means. Once AIs can crawl through contracting and legal paperwork and capture relevant information, intelligence operations will be able to put together a very complete picture of where and how adversaries are spending their money.
Tactical Auto-Aim	2	1	4	Doubles the power of your force tokens. (If no VVTE was conducted, the chance of IP loss and amount of loss are both doubled)	When guns detect and shoot at targets in a millisecond, overwatch replaces suppressive fire, and untrained personnel becomes sharpshooters.
AI Enhanced Propaganda	2	0	7	Grants the ability to spend (1) per COCOM to triple total military presence for this turn.	Convincing local governments and populations that your military is powerful can be done by having a powerful military... or by some exactly targeted press coverage and social media posts.
AI targeted Recruiting	2	0	4	Develop 1 free force token a turn. Lose 1 IP a turn. (This technology causes twice as much IP loss if adoption fails)	The difference between 'creative recruiting strategy' and 'poaching' starts to blur when algorithms can reach individuals with tailored advertisement messages.
Automated Cloud Environments	1,2,4	0	9	Reduces the cost of adopting all AI techs by 2 (min cost of 1). Increase RP gained per turn by +1 for the rest of the game.	Cloud services allow automation of almost everything except haircuts. If the initial costs are paid and the environment is set up right, any work not requiring creativity or extremely advanced decision-making can be eliminated, simplifying jobs across the entire force.
Big Data analysis	5,3	0	3	Displays all other players' current IP scores.	AI technology allows Intelligence analysts to actually USE all of the massive amounts of data they can collect, instead of cherry-picking based off of intuition and simple

					heuristics.
Deep Fakes	2,5	0	4	Once a turn, gives you the ability to freely craft the player, effect, and location of an objective for the next turn. Can be used for all types of objectives, and can be used on yourself.	Deep Fakes can allow spoofing of communication in the most trusted medium right now: video. Used externally, it can cause other militaries to chase their own tail. Used internally, it can influence elections, policy decisions, and the opinions of entire populations.
Fully Agile Development	1,2,4	0	3	Reduces cost of adopting all AI techs by 1	Agile software development, while not directly related to AI, is almost a necessity if an organization wants to be 'AI-Ready.' It's been the standard commercially for many years now.
Satellite Image Analysis	4,5	0	4	Each turn, reveals the projected moves of random(0-8) enemy forces. For (2) RP, you can permanently increase the number of revealed forces by 2.	Augmented by high-fidelity satellite images and video, advanced AI systems can make accurate predictions for upcoming enemy force movements.

Table 3: Tech Card effects, including related LOs, how early they can appear in the game, their cost, the game effect, and the flavor text displayed to players

Development of AI Opponents

One of the most compelling aspects of a game is the competition. SGs designed to educate players are no exception to this rule. To that end, the AI opponents for Obsolescence were designed to only allow human players to win if they understood both the game's mechanics and its intended lessons. AI opponents were designed as reactive behavior-based agents utilizing a set of pre-computed weight tables. The weight tables were populated through a simple reinforcement learning approach and the arbitration between low-level behaviors was conducted from a combination of the weight tables and the game states. The final model was tested against variants with un-trained weights, random weights, and against an older game AI.

At its core, every behavior exhibited by the AI is a stand-alone module, capable of taking a certain set of actions within the game to achieve an effect. By moderating between the different behaviors, the AI can make decisions that optimize towards higher IP gain. As a reactive system, this AI does not utilize look-ahead mechanics and only reacts to the current state of the game [44]. It uses three sets of weights, two arrays, and one dictionary, to change behavior as the game progresses. While those weights reflect predictions upon the future game state, those predictions are not created based on any input the AI is receiving, nor are they modified in any way as the AI runs. Instead, the weights used in the final model were created in the training phase. Figure 7 describes the decision tree utilized by the AI.

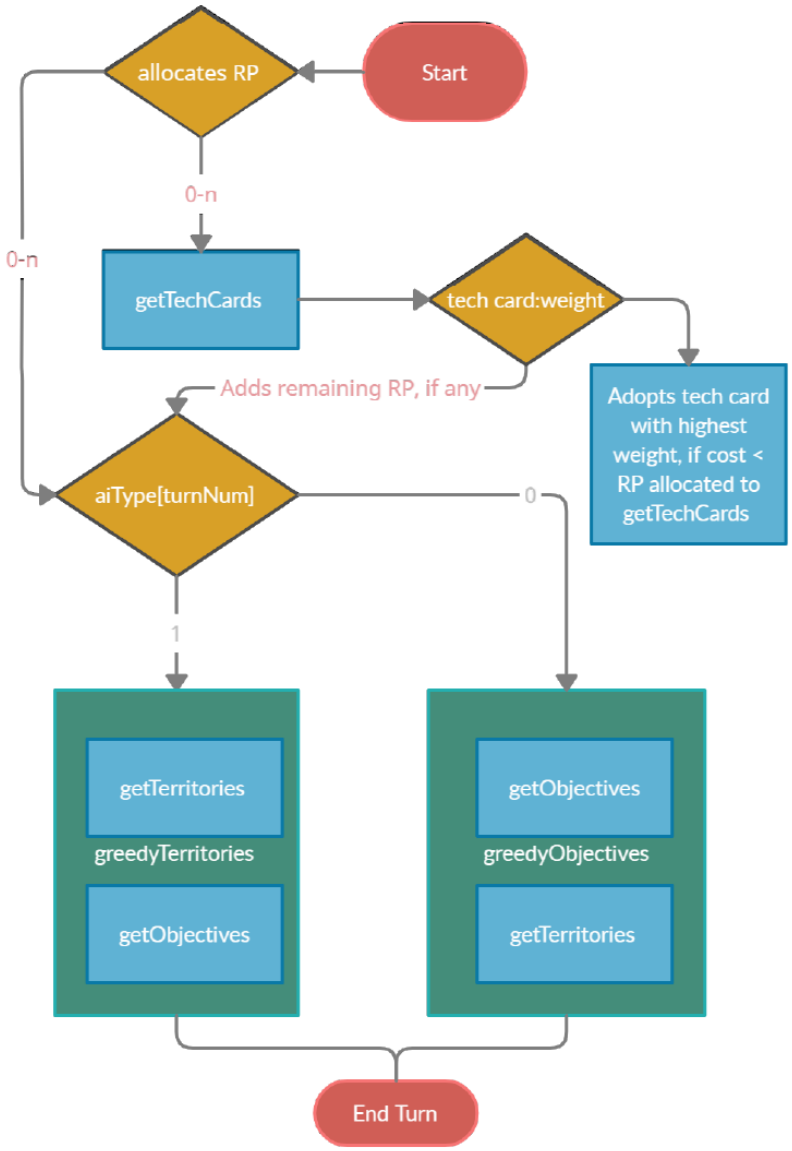


Figure 7: The AI's decision tree for each turn's actions

There are three low-level behaviors the AI uses to interact with the game, *getTechCards*, *getTerritories*, and *getObjectives*. The latter two are combined into two separate behaviors, *greedyTerritories*, and *greedyObjectives* using a fusion of the low-

level behaviors. In total, the behaviors operate on the set of inputs and produce outputs (Table 4).

Possible Inputs	Possible Outputs
Board State: Token Placements, Current Objectives, Available/purchased Tech Cards	Moves force tokens around the board
Current Turn number	Builds new force tokens
Weight Tables: TechCardWeight, aiType, RPSplit	Adopts new Tech Cards

Table 4: Possible Inputs and outputs to AI decision-making

The weight tables were generated using a simple training method. The weights were determined via a cycle of training games. Each iteration of the training cycle ran the AI against itself. After 10 games, the unique set of weights that won with the highest score became the new baseline weight set. Agents for the next 10 games slightly adjusted the weight ratio from the baseline weights by having a 33% chance to modify each value in the arrays by either +1 or -1.

This cycle ran for three sets of 2,000 games and was repeated 8 times with independent starting values. The 8 most successful sets of weights were then ran against each other for another 3,000 games. The weight set displayed in Table 5 was the set with the highest win rate.

CardName	AiRadicalization	AiTesting	AutomatedCloud	AutoAim	RoboLogistics	AIRecruiting	BigDataAnalysis
CardWeight	6	4	5	4	5	5	5

Turn	1	2	3	4	5	6	7	8	9	10
RPSplit	2	1	2	2	3	2	1	1	3	0
AIType	0	0	0	1	1	1	1	0	1	0

Table 5: Final weight set for Obsolescence's in-game AI opponents

To verify the process, the final weight set was evaluated against random weights, no weights, and an older version of the AI that did not purchase any Tech Cards at all. It outperformed them all, winning an average of 62% of the games. Concluding this process, the final weight set was permanently added to the AI algorithm in Obsolescence. Anecdotal testing shows that it performs strongly against human opponents. The game developers and two volunteers reported a higher challenge when facing off against the new AI.

Relations to Educational Goals

The game mechanics for Obsolescence were designed to engage with each of the LOs. The following sections describe how each LO influenced the game design, and which game mechanics satisfy the educational goals of the LO.

LO1. Defend the value of VVT&E for all DAI technologies

Before adopting a Tech Card, players can *conduct VVTE* for each card. This replicates real-world project management decisions and allows players to learn the potential benefits and downfalls of VVTE through repeated decision-making. Adopting a technology without thoroughly testing it and assigning a proper usage for it can cause slow-down, waste, or ethical catastrophes.

From a gameplay perspective, players have the option to start adopting a Tech Card as soon as it becomes feasible. However if they don't take a turn to properly evaluate it, the adoption may fail. Some cards have more significant effects than just loss

of resources and time if a player does not conduct VVTE. For instance, auto-aiming weapons carry an additional possibility of a significant IP. One of the worries with a lot of autonomous weapons systems is the problem of blue-on-blue or blue-on-green fires. An event such as that not only costs time and resources but can have significant international and internal fallouts. This is represented in the game by a deduction of IP, abstracting the myriad complicated detrimental effects into the game's victory condition.

LO2. Recognize that DAI can/will greatly increase the complexity of the military environment.

There are three ways in which the game mechanics are designed to teach LO3. First off, several Tech Cards are additive. They increase the amount of game mechanics occurring in a given turn, making it harder for a player to accurately grasp and predict what the current or future turns will look like. With specific card combinations, a player can have unlimited moves available and infinite resources.

In addition, new Tech Cards becoming available each turn increases the player's information. At Turn 1, the player has three resource points to allocate to an average of 8 potential moves. (5 force movement locations, 1 build force option, 1 technology VVTE, and 1 technology adoption). At the end of the game, Turn 14, the player will have significantly more forces, which are no longer homogenous, and each COCOM will have a vastly different makeup of forces in it. Each force token has 5 possible moves. Now, there are up to 15 technologies to adopt or VVTE, and up to 65 new choices that can be made from adopted technologies. Effects from Tech Cards, both those adopted by the player and by the opponents, will require prediction and calculation changes from turn to

turn. The Tech Cards also add variation between games, making no two games identical. Heuristics that worked as a strategy in one game may need to be adjusted in the next to deal with different emerging technologies. These game mechanics create an environment where players can quickly get overwhelmed with the options and information becoming available each turn.

LO3. Support and value increases to military 'Complexity Carrying Capacity'

For this LO, several AI technologies were added to the game specifically to aid players with decision-making. These cards do not give an in-game advantage directly, but give the player more time, more information, or clearer strategies to combat the increase in complexity. Players who adopt those technologies can make more informed decisions, offsetting the cost of technology adoption. The game is designed to be significantly harder without using those technologies, reinforcing the idea that increased complexity carrying capacity is vital for military success.

LO4. Assess value of strategic plans and roadmaps that deal with DAI

As a turn-based strategy game, the format lends itself to planning turns in advance. Technologies get cheaper to adopt over time and players can save resources to achieve more resource-intensive goals. In addition, repeated playthroughs of the game give increased familiarity with the potential Tech Cards that may appear. This allows players to make strategies based on potential technologies, both for their plans and for planning around their opponents. In doing so, players will critique and refine their strategies, developing skills for evaluating real-world proposals in similar domains.

LO5. Support and value increases to international monitoring and restrictions on AI progress and development

To win, players either need to monitor their opponents or get extremely lucky. The game was designed with a clear 'winner' and 'losers' in mind, not for realism, but to encourage direct competition between players. While there are no in-game options to enforce the equivalent of international technology restrictions, the game has multiple settings that can be configured. For example, players can choose to play a game with fewer available technologies. Changing the game settings can easily make the game more manageable for human players, much like international treaties can allow two militaries to have a humane and contained conflict. Players who utilize these setting changes, or who can postulate theoretical changes to the game, may be able to see the rationale behind international restrictions on AI progress.

Relations to Research Questions

Infrastructure design

The game was designed to be playable from commercial devices, including smartphones and laptops. The code for the game is entirely JavaScript, which enables the game to run on most modern web browsers. This design decision allows Federal employees to participate from both work computers where an executable file would be blocked and from home computers where computing power or hard drive space may be limited. The game is hosted on a Microsoft Azure compute instance owned by AFRL's Hanger18, which enables global distribution and scalability if required. Hosting is

enabled by the containerization of the web server and game files. In total, this enables participants to start play-testing the game with only a single link, and administrators to automate the entire pipeline from the developer's workstation to the production website.

RQ1: Does the game Obsolescence teach its Learning Objectives?

To best support answering RQ1, the game design included a logging system for in-game actions taken by human players. As explained in chapter 3, these logs can be used for educational analysis, especially when correlated with the surveys players take. Logs capture the following information: playerID; time spent in-app; time spent reviewing LOs; (for each game-)Total IP; time spent in-game; (for each turn-)Techs adopted; techs VVTEd; IP; Force moves made; adoption failures

Some aspects of these logs may correlate to specific behaviors demonstrated by players with high levels of learning. When analyzed at a sufficient scale, the logs may also reveal interesting trends that indicate learning being expressed through certain game actions.

RQ2: How does the measurement of learning compare to the reported learning?

Several of the game logs also assist researchers in investigating RQ2. Specifically, logs for total time in-app, time in each game, and total count of games are tracked for later correlation with participants who reported learning. The number of games played can be correlated against the reported engagement and enjoyability of Obsolescence. All the game logs have the potential to correlate to the reported and

calculated measures of learning but would require a large data set to be considered statistically significant.

Conclusion

Obsolescence was designed and built for this experiment, and to provide a potentially valuable educational tool in an area relevant to the DoD's current interests. As such, the game was designed around five LOs, stemming from recommendations and overall guidance from DoD think-tanks. The structure of the gameplay was based on real-world observations and from the RAND tabletop game *Hedgemony* [11]. Instead of human opponents and a Game Master, Obsolescence used a custom AI opponent. Each of the Tech Cards and much of the gameplay itself was designed to support the 5 LOs. The game's design also took into consideration the two RQs and the process of conducting an online experiment.

IV. Methodology

Chapter Overview

This chapter details the process of experimental evaluation of the digital SG Obsolescence. The purpose of this research includes addressing three factors: to what extent the game teaches participants the LOs, the amount of provided engagement and entertainment value, and an analysis of reported metrics vs measured results.

The study used two online surveys and direct measurements of in-game logs. The questions used for the survey were pulled from authoritative sources on the game topics and were weighted based on expert opinions. Participants take one survey before and one after playing the game. All survey questions were measured on a 5-point Likert scale. Analysis was conducted utilizing differences between the pre- and post-surveys, averages and standard deviations of post-survey questions, and correlations between the direct measurements taken by the game and the corresponding survey answers.

This Chapter has two sections: the generalized methodology behind game assessment, and the specific experimental methodology for evaluating Obsolescence. The description of the game design methodology is in Chapter 3.

Game Assessment Methodology

This section describes a novel educational game assessment methodology. This methodology gives researchers a tool to more objectively evaluate the success of an educational game broken down by individual LO. It also assists researchers in measuring

educational serious games that teach LOs measured on Bloom's affective taxonomy [10], [21]. This methodology is distinct as it provides a generalizable framework for game assessment that mitigates subjectivity from self-reporting, builds a baseline to measure against, gives a standardized format for tracking each LO, and allows for testing knowledge captured in higher levels of learning taxonomy. Figure 8 provides a graphical summary of the entire process and acts as a one-page handout to promote the assessment methodology.

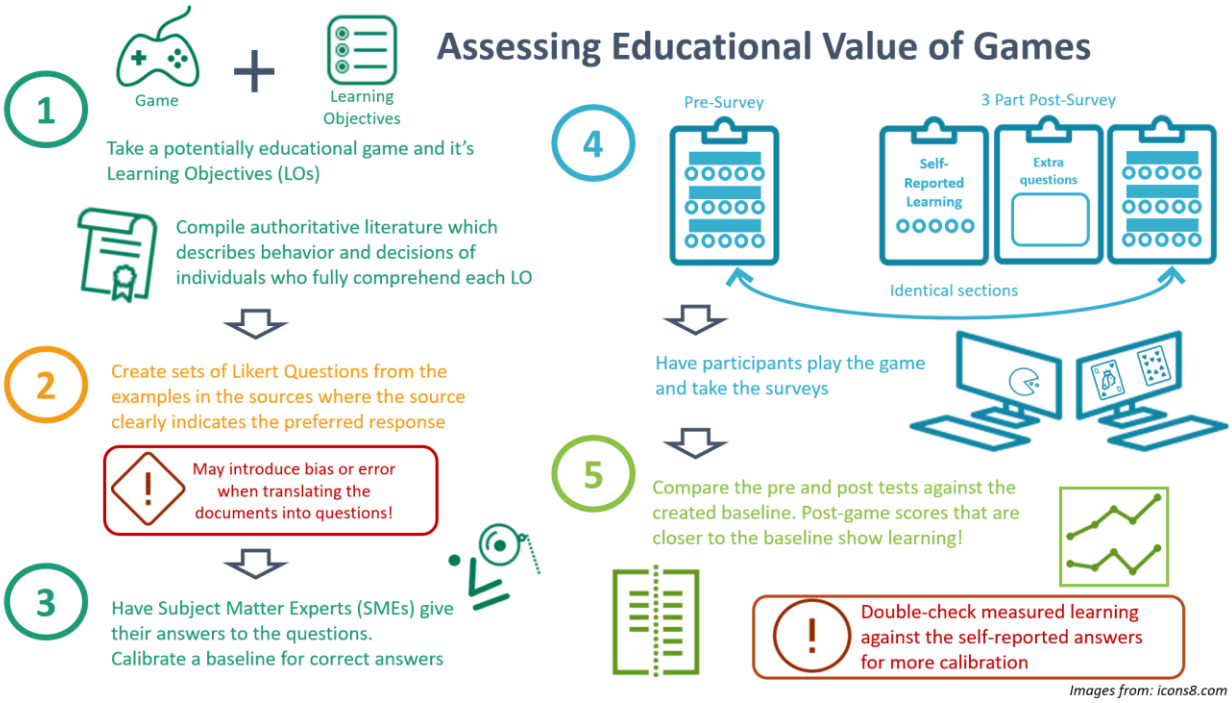


Figure 8: Methodology handout describing the assessment's procedural flow

1) Identify desired learning objectives.

The first step for evaluating an educational game is to define the desired LOs. This methodology is appropriate for LOs that cover complex or not well-understood

topics. Such topics often fall under the affective domain or high levels of the cognitive domain [10]. Researchers should also compile literature that describes exemplar behavior for individuals who fully understand each of the LOs. For example, Obsolescence used reports that recommend specific actions to conduct more and better VVT&E for AI technologies. Those actions can be confidently said to be the actions of an individual who fully understands Obsolescence's LO1: Defend the value of VVT&E for all disruptive AI technologies.

2) Capture the behaviors students should learn in the format of Likert survey questions.

For each of the LOs, a set of related questions should be crafted. Each question should assess a facet of the LO and the set of questions should sufficiently address the intent of the LO. Questions relating to higher level LOs may be opinion questions, without an objectively best answer. While each question might be answered incorrectly by a participant who truly has learned the material, having a series of questions all correlated to the overall LO increases confidence in the overall assessment.

To mitigate researcher bias, questions should be sourced from the official or authoritative sources researched in step 1. Textbooks describing the high-level LO may often give examples of behaviors exhibited by individuals with a strong understanding of the topic.

This is the step with the most likelihood for error. Reducing a complex topic into a set of survey questions requires arbitration from the researcher, and will likely create

information loss. The final set of questions and scores for each LO may not capture the essence of the LO or evaluate the full range of learning for a complex or subjective topic. Therefore, researchers should derive their questions as directly as possible from external, authoritative sources to ground their questions in previously established scenarios.

Example:

To use an example from Obsolescence, LO2 is "Recognize that DAI can/will greatly increase the complexity of the military environment." After explaining and stating the above LO, NSCAI provided recommendations of actions the DoD should take [14]. These actions showcase individual decisions that are heavily influenced by a strong understanding of LO2.

- "We recommend the DoD divests from military systems that are ill-equipped for AI-enabled warfare, instead investing in next-gen capabilities"

- "We recommend the DoD assign an AI Operational Advocate on the staff of every Combatant Command. This officer would perform a similar role to that played by the Staff Judge Advocate. He or she would be an expert in AI systems, advise the commander and staff on the capabilities and limitations of AI systems, and identify when AI-enabled systems are being used inappropriately."

- "We recommend the DoD Integrates AI into major wargames and exercises to promote field-to-learn approaches to technology adoption."

For the evaluation of Obsolescence, these recommendations were turned into the following questions, with possible answers ranging from (1) S*trongly Disagree* to (5) *Strongly Agree* for Likert-scale analysis:

- "The DoD should cease development and funding for military systems that are ill-equipped for AI-enabled warfare."
- "Every COCOM staff should add a new member (similar to the staff JAG) that is an expert exclusively on AI systems."
- "AI systems, applications, and scenarios need to be integrated into all major exercises."

For the above three questions, the initial intent was for participants to answer a score of (4), (5), (5). The wording of the first question was stronger than the original NSCAI report, and the original report more closely aligns with *Agree* than *Strongly Agree* [14].

When determining the phrasing of the question, questions can be designed so that the baseline score is not an extreme (*Strongly Agree*/*Strongly Disagree*). This encourages refinement of the grading system in step 3, as expert opinion can bump the baseline score up or down.

3) Have experts take the survey questions to establish baseline scores

The transformation from example behaviors or theoretical actions will inevitably introduce some drift from the original intent. To assist in the calibration of the baseline, subject matter experts (SMEs) should answer all the questions. Their responses are used

to calculate the baseline metric for evaluating if students have achieved the desired learning.

The process of utilizing experts can also be conducted and repeated to aid the design of the questions. For instance, if the experts do not agree on a particular score to a question, the question should be reworded or removed.

It is important to select sources and questions that apply to this educational objective and avoid basing survey questions on sources that might be overly specific, or whose answer relies heavily on context. The expert baseline helps mitigate those effects, but, ultimately, might itself suffer from similar issues. This could occur when individual experts disagree based on their field or local context.

Expert opinions should be weighed against the original sources' intent, at the researcher's discretion. For this research, the baseline score for each question was calculated as shown in Equation 1 by averaging the expert scores and the original intent of the source material from which each question was derived. The particular formula used for this experiment is arbitrary and would likely change with different sources for questions and expert populations.

$$baseline\ score = \frac{2 * (average\ expert\ score) + intended\ score}{2}$$

Equation 1: Baseline score calculation, applied to each question

The SME calibration does not guarantee that any question adequately captures the correct learning. Creating a baseline in this manner partially replaces the need for a

control group. Using external sources and experts, instead of just the researchers' knowledge, helps measure more subjective LOs without creating an assessment of the game that heavily relies on the evaluator's knowledge and preferences.

4) Create pre/post-survey questions and perform the experiment

The pre-survey consists only of the questions created in the above process. The post-survey is composed of three parts. To avoid any experience during the survey portion of the experiment affecting the self-reported metrics, the scenario questions created to measure learning are administered last. The specific surveys used for Obsolescence can be found in Appendix B.

Part 1- Direct questions: "Did you learn the LOs?"

These questions are standard for many current evaluations of games or other educational material. This type of question directly answers the educational goals but relies on the participants' honest and accurate self-assessment. These questions may be subject to participant bias and may not capture learning that the participant has not themselves realized. This problem becomes significant when researchers attempt to measure more abstract and/or higher-level learning objectives.

Another method of avoiding personal bias and self-knowledge is to create questions about other participants. These questions would be applicable in group learning experiences if the game was multiplayer or team-based. Participants would be asked if another individual demonstrably achieved the LO, and their responses can be used to offset the participant's self-assessment.

Part 2- Engagement evaluation and short answers.

The most commonly reported advantage of game-based learning is increased engagement [18]. A complete evaluation of the game should therefore also assess how well it functions as a game, not just as an educational tool [29]. Questions on ease of use and frustration with the hosting infrastructure are also appropriate here.

This section can also include short answers to other questions potentially of interest to the research. If the game can be modified or is in development, this is also where researchers should add questions related to game development and game design.

Part 3- Identical survey questions to the pre-survey

In the last section, the post-survey will ask identical questions to the pre-survey. Participants will have had no experiences other than those playing the game. To account for pre-game knowledge, these questions are only relevant when compared to the participants' pre-survey. To that end, it is critical to assign a control number to each participant and attach it to both of the surveys for future analysis. Participants should take both surveys directly before and after playing the game, to ensure the surveys are measuring only the effects of the gameplay.

5) Analyze for Learning and Engagement

Part One and Two of the post-survey ask participants direct questions about the game pertaining to the game's educational effectiveness and their level of engagement. Both parts can be analyzed using simple statistical techniques, such as identifying the

mean score and its standard deviation. Results from this analysis should be careful to mention that these answers are all self-reported measurements.

The pre-survey data is only useful for game evaluation when paired with the post-survey Part Three. Researchers should analyze any delta between pre and post-surveys to see if students have changed their opinions, views, or knowledge. The magnitude of the delta suggests evidence of learning, while any shift towards baseline scores measures the satisfaction of the LO.

Researchers can analyze participants to see how many, if any, modified their scores to more closely resemble the established baseline. If this is common among participants, this would signify that the game is teaching the LOs. Participant scores should be evaluated on how close to the baseline their responses were. Participants skipping a question does not discount the question from the analysis; on the contrary, it may indicate that a participant did not feel confident giving any answer. Any response on the post-survey would indicate that they now feel more informed about the topic.

Lastly, the results from comparing the pre-survey and post-survey Part Three can be contrasted with the direct questions in the post-survey parts one and two. Ideally, participants who confirmed they found the game educational would also demonstrate their learning by a change in their pre/post responses. The combination of both data sources can help mitigate both the bias incurred by the direct questions in parts one and two and can mitigate the indirect nature of the questions asked in Part Three.

6) Analyze Direct Measurements of Game Logs (Optional)

This step may not be possible to complete based on the specific game being evaluated. For the study of Obsolescence, the game was created in-house and the researchers had full access to the source code during and after development. The needs of the experiment heavily drove the development of the game, as outlined in section 4: Design of Environment. However, the experiment's population was small, limiting the usage of the game logs.

The logs collected from the educational game need to have a control number linked to them, so researchers can correlate the gameplay with the surveys. If possible, analysis of in-game actions of the participants that demonstrated the highest level of learning can greatly assist future usages of the game. A strong enough correlation may allow instructors to evaluate future students' learning using in-game metrics instead of surveys. For instance, if the participants who learned the most all eventually used the same strategy, 'a utilization of strategy X' could be used to evaluate when players have achieved the desired learning. Instructors who use this serious game could therefore make it more accessible by removing the surveys and using only in-game metrics. Logs can also be used to determine the optimal time spent in-game to achieve measurable learning. If the participants with measured learning also reported similar times spent in-game, that amount of time can be implemented to game-play by educators.

In-game logs, as direct measurements, are extremely useful for analyzing usability and enjoyment measures. Participants are asked usability and entertainment questions in the post-survey, however that data will be undoubtedly biased. Participants' perceptions

of events may differ from the actual occurrences. For instance, direct measurements of time spent in the game can be correlated with reported time spent in-game. Participants who overestimate the time they spent in-game may have found the game boring, whereas participants who reported less time than they spent may have genuinely enjoyed the game.

This step can also be extremely helpful when designing a game or the educational program utilizing a game. Direct feedback from game logs can indicate which areas participants are spending the most time in, or which aspects of the game are needed to reinforce the LOs. For instance, the Tech Cards in Obsolescence were each designed to help teach one or more of the LOs. Direct reports of game data could help developers balance the game to ensure there exist viable (and enjoyable) strategies involving usages of every LO's Tech Cards.

Experiment Design for Obsolescence

Subject Recruitment Plan

To recruit subjects for playtesting Obsolescence and the game evaluation methodology, the following steps were taken. First, both the game and the survey questions were approved through both AFIT's/AFRL's Institutional Review Board (IRB) and Public Affairs. Participants were recruited via a combination of an email campaign and typical channels such as Air Universities Microsoft Teams. Table 6 lists the federal organizations that had access to Obsolescence.

Educational Centers/Courses	Military Units	Other
School of Advanced Air and Space Strategies (SAASS)	LeMay center for Wargaming	Contact list for current and retired DoD wargamers
Air University (AU) Teaching and Learning Center	Joint AI Center (JAIC)	AFIT student population
USAF Air War college AI/ML elective	AFRL Trusted Autonomy, Cyber, and Serious Games	AFIT cyber operations track
US Marine Corps University (USMCU)	USSOCOM's AI Portfolio Management Office	
US Coast Guard University (USCGA)	Office of the DoD Chief Data Officer (CDO)	
Naval Post Graduate School (NPS)	711th human performance wing	
AFIT Cyber 101	88th Communications Squadron	
North Dakota State ROTC		
AFIT Intro to Autonomy		

Table 6: The organizations given access to Obsolescence

As per the research protocols, the subject population was limited to Federally employed individuals, and all research activities were completely voluntary. No reward was given for taking the surveys, and no expectations were levied upon personnel from their supervisors/chain of command. The experiment's website was controlled with a simple authentication policy and ran during the dates 16 November 2021 - 16 January 2022.

Experts were sourced from directly contacting authors of the sources used for this paper, and from identifying individuals in the participating organizations who worked in either the field of military AI, AI technology prediction, or military strategic planning and who self-identified as experts.

A short pilot study was conducted one to two months before the experiment itself. This study asked a small population size (n=7) to play the game and report feedback on the game design and effectiveness as a teaching tool. It also asked a larger population (n=28) to take the pre-survey and identify potential issues with the wording of the questions. The feedback on the game prompted several User Interface (UI) design changes and an update to the tutorial. The pre-survey questions used in the pilot study were phrased too positively. Participants reported extremely high scores across almost all questions. The questions were subsequently reworded to be more extreme to provide more opportunity for answers to shift after playing Obsolescence. When the new questions were given to a portion of the pilot study population the distribution of scores was larger and the average score was lower. The pilot study also confirmed that the data collection methodology worked as designed.

Experimental Procedure Steps

Participants were asked to participate as follows:

- Receive login information, including username/password and Informed Consent Disclosure.
- Receive a control number from the game
- Take the pre-survey questionnaire

Participants may participate in any combination of the following:

- Review the in-game Tutorial

- Play games against AI opponents
- Adjust game settings (number of players, game time, starting resources, etc)
- View more information on the technologies/game concepts
- Exit the game, whereby they are invited to fill out the post-survey

The game provides the post-survey link and attaches the control number and in-game logs to the post-survey data. The surveys were conducted using Google Forms and did not collect any PII information, including email addresses or IPs.

Data Analysis Plan

Research Questions:

RQ1: Does the game Obsolescence teach its Learning Objectives?

RQ2: How does the measurement of learning compare to the reported learning?

Analysis of post-survey questions directly asking about achieved LOs and Engagement.

This analysis assists with both RQ1 and RQ2. For Obsolescence, these questions will be analyzed in a parametric manner. The questions requiring short or long answers will be individually analyzed by researchers. If participants do not answer the questions about time spent in-game, the data pulled directly from the game logs will be substituted.

The experiment will track the distributions of answers, both the mean, standard deviation, and outliers.

Analysis of the delta between post and pre-survey questions.

This analysis assists with both RQ1 and RQ2. Participant data is organized using a table of scores for their pre- and post-surveys. These will be compared against each other and the established baseline. This comparison will generate data on the degree to which participant answers changed either towards or away from the baseline. If a participant answered the same on both tests, their score for that question is 0. If they answered closer to the baseline in their post-test, their score is a number equal to the numerical value of the difference, positive if they moved towards the baseline, and negative if they moved further away.

This mitigates the potential disparity in knowledge participants may have before coming into the experiment. If a participant scores each post-survey question with exactly the baseline scores, this only indicates they learned their knowledge from Obsolescence if their scores on the pre-survey were far from the baseline. Otherwise, this particular individual likely already had a strong understanding of the LOs and the game did not teach them anything significant.

The results from these comparisons will be analyzed both in aggregate and on a per-LO basis. For each, the research will identify the mean and standard deviation of the total change towards the baseline that participants demonstrate.

Analysis of any correlations between the measured and reported learning.

This analysis assists primarily with RQ2. Using reported levels of learning from Part One of the post-test, a correlational analysis will be conducted between the measured learning and the reported learning. This analysis will include calculations for statistical significance and will be conducted both in aggregate and for each LO. A significant correlation indicates evidence that both the reported scores and the measured scores are studying the same phenomenon. If both the scores indicate a positive learning experience, Obsolescence will have demonstrated educational potential. A lack of correlation could indicate one or both of the measurements failing to accurately capture the game's value, or may hint at methodology problems with either the reported or measured metrics. For instance, participants that score the game's educational value highly only in an attempt to be nice to the researchers would not have a correlated measurement of learning. Alternately, the questions built to measure the learning may not be sufficient to differentiate between participants who truly learned the game's LOs and those that did not.

Conclusion

This chapter outlines a generalizable methodology for assessing the educational value of a game. The methodology was designed to overcome some of the common scientific shortfalls many educational assessments face and to give the ability to measure opinion-based questions. As a novel methodology, it also includes standard survey questions for participants to self-report their learning, both as a backup assessment tool

and a calibration tool to confirm the merit of the assessment methodology. This chapter also contains the specific steps used by this research to follow the methodology in conducting its experiment.

V. Results and Analysis

Chapter Overview

This chapter describes the results and analysis from studying player learning after playing the SG Obsolescence. This experiment was hosted using a public-facing website that did not log connection information. The credentials for access were distributed across 19 DoD organizations, potentially reaching thousands of individuals. Data was collected from Nov 3rd to Jan 12th and consisted only of the data provided from the surveys. Of those that accessed the website, 48 participants submitted the pre-survey form, 31 submitted the post-survey form, and four SMEs gave their opinions on the pre-survey questionnaire. Of those participants, 24 submitted both surveys with the same control number. The pre-survey responses that were neither expert opinions nor correlated with any game logs or post-survey responses were not analyzed. In addition, game logs were obtained from 76 game playthroughs.

Data Preparation

Several data collection and reporting issues may have influenced results. These were identified either by participants informing the researchers or identified by the researchers after the experiment window had opened: non-contiguous game-play or alternate survey access, inaccurate game log data, and verbiage change in the surveys during the experiment.

Users who closed and later reopened the game would not only fail to submit their game logs from the earlier session but would also reset their control number. This likely was the cause of many of the post-survey results that did not have a matching pre-survey submission. In addition, users who accessed the game on their phone or tablet were able to play the game entirely but had issues accessing the Google form links. The extent of these issues is unquantified but is expected to be relatively low. Some participants may have generated fewer data points in the post-survey than occurred in-game. This would occur if the player opened the post-survey link before completing their game playthrough.

The logged data from the game also held inaccuracies. Game logs correctly tracked the technology cards adopted by each player but did not log the turns each player adopted the technology. Game logs for gameplays where the participants exited back to the main menu without completing the game were not recorded properly, and could not be used.

Lastly, the text on the pre-survey form was modified slightly a few days after opening the experiment by replacing every instance of "VVTE" with "Validation, Verification, Testing, and Evaluation (VVTE)" after the request of several participants.

Establishing the Baseline

The baseline was developed following the procedures outlined in Chapter 4. All the questions created for this experiment had an associated score appropriate to the original intent of the source material. The researchers' generated this score by

interpreting the original intent of the source material with regards to the five-point question. To mitigate the subjectivity created by such interpretation, four SMEs took the survey questions and reported their answers for each of the questions. The equation for calculating the baseline can be found in Chapter 4, and the results from the calculation can be seen in Table 7.

Question Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Intended Score	4	5	5	4	5	5	4	4	4	5	5	5	5	5	5	4	4	4	5	5	5	5
Average Expert Scores	3.5	4.5	4.3	3.8	4.5	4	4	3.8	4.3	4	4.5	3.3	4.5	3.8	3.8	3.5	4	4.3	4.5	4	5	4.8
Calculated Baseline	4	5	5	4	5	4	4	4	4	4	5	4	5	4	4	4	4	4	5	4	5	5

Table 7: Baseline generation using experts and intended score to calculate a baseline score for each question

In this case, as the questions were sourced from positive recommended actions for the DoD, the intended scores were all either (4) *agree* or (5) *strongly agree*. Most experts' opinions were consistent with the intended score of most of the questions. However, one of the experts gave significantly lower scores for many of the questions. While the expert scoring was conducted anonymously, from discussions with several SMEs, this is likely due to a personal belief that the military should entirely refrain from competing with AI technology. This reveals a potential for error when measuring learning. Participants who hold similar contrarian perspectives may learn the values and skills taught for each LO, but interpret them in an unintended and unmeasured fashion. None of the SMEs identified other perspectives that would lead to a participant who experienced learning modifying their answers away from the baseline.

Reported Learning

The first research question asks to what extent the game teaches its LOs. The study directly asked participants questions related to RQ1, collecting self-reported statistics for each LO. Figure 9 describes the reported learning from the post-survey part 1. Each participant was asked five questions, one for each specific LO. Each box displays the standard range of answers for that question/LO and indicates the mean response. Recall from Chapter 3, each survey question provided a Likert scale where (1) is *Strongly Disagree*, and (5) is *Strongly Agree*. Overall, the average response across all five questions was 3.8, *Agree*.

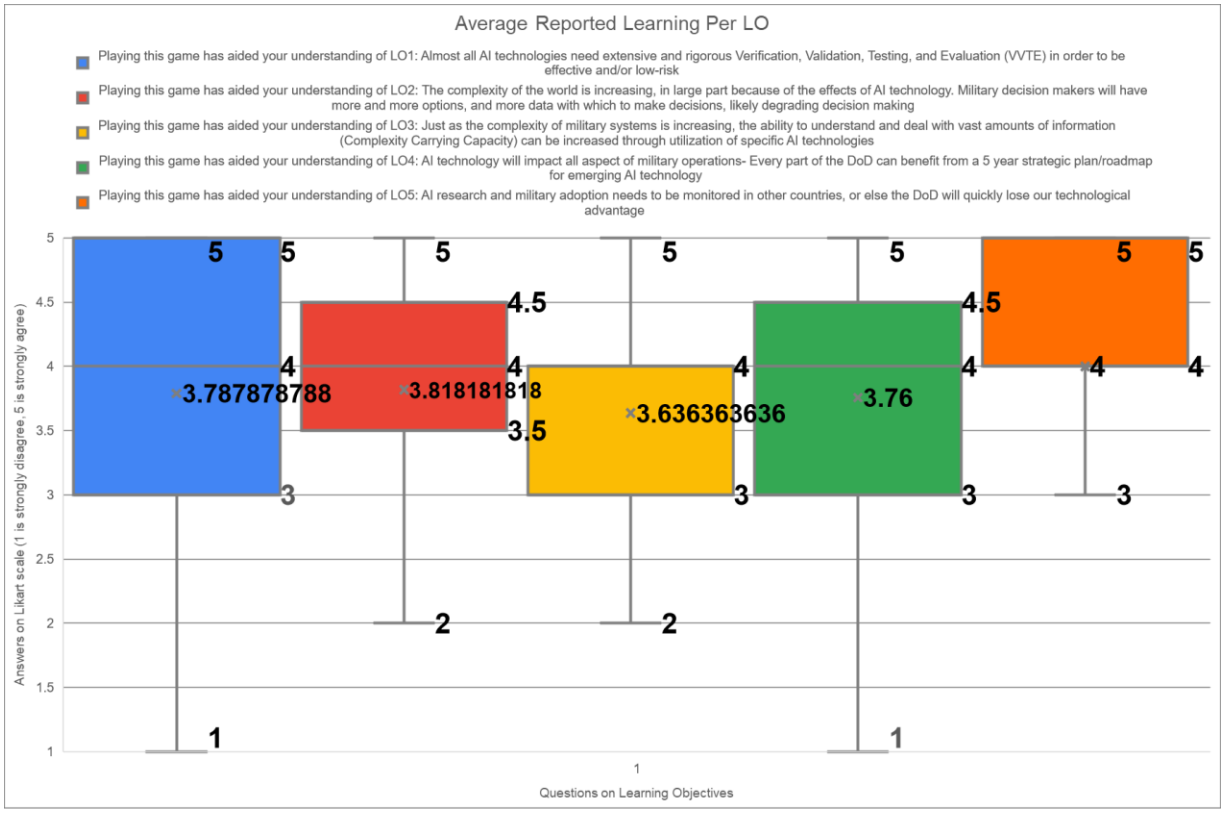


Figure 9: Reported Learning for each LO

Part 2 of the post-survey provided an opportunity for participants to give short answers. The most frequent complaints, in response to short answers on enjoyability and educational benefit, were about interface or tutorial frustrations. Players experienced frustrations such as "unreadable text," "not clear why I start the turn with 5 or 7 RP," or "clarify the ground unit interactions." This may explain why participants rated LO3 the lowest, as they did not experience technology aiding their ability to understand the complexity of the game. The players also reported wishing for an improved tutorial and suggested making the tutorial mandatory. The players' information processing and decision-making abilities are supposed to be assisted by cards and game mechanics; however, if the interface is degraded or players do not understand parts of the game those helpful Tech Cards and mechanics may not have been able to work as intended.

Conversely, LOs that relied on game mechanics or the nature of the game itself scored higher on the post-survey. LO5 was taught mostly through the nature of a competitive hidden-information game. Several comments complained of the lack of AI technologies designed to help with LO5. Players recognized the need for such technologies and wished for Tech Cards with additional abilities for increased adversary observation.

Measured Learning

In addition to the reported learning, this experiment answered RQ1 based on the novel assessment methodology outlined in Chapter 4. Comparing the change in participant answers from the pre-survey and Part Three of the post-survey indicates the

mindset change and the learning after playing Obsolescence. Participant scores were more similar to the baseline after playing Obsolescence.

Figure 10 shows the average change per LO and the total change. The data indicate that playing Obsolescence has motivated participants to change their values closer to those of the authoritative sources and the SMEs for all but LO4.

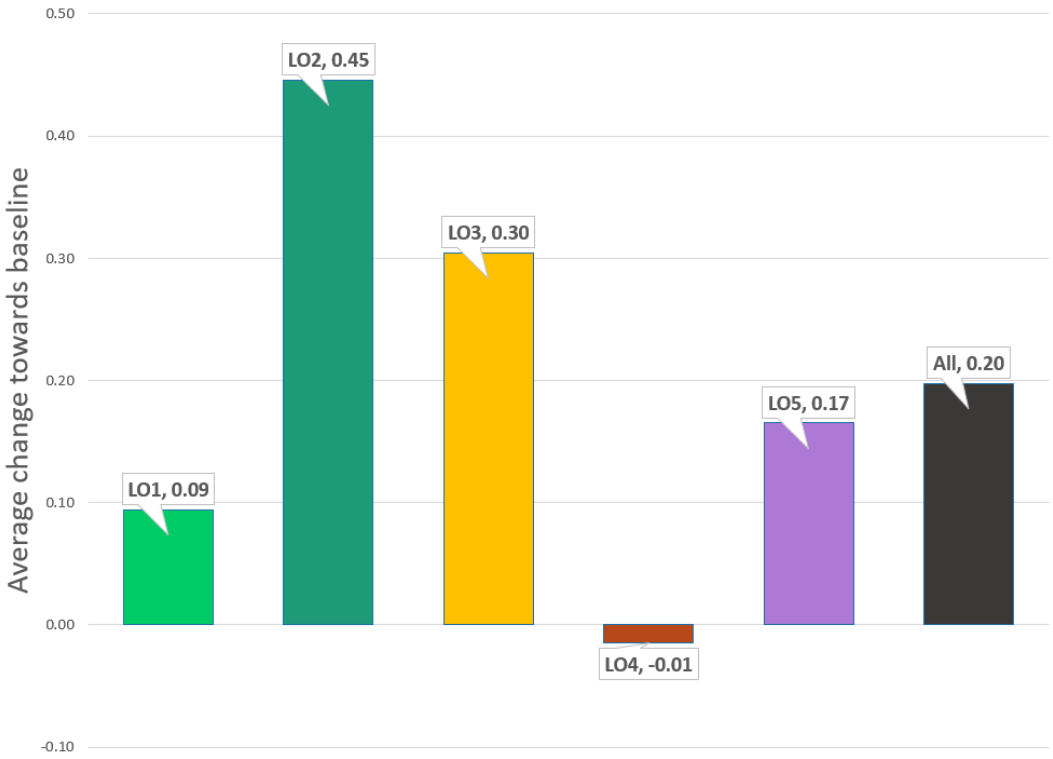


Figure 10: Average change towards baseline per LO

Figure 11 displays a comparison between the participant's pre- and post-surveys and the baseline scores. The graph charts the average change in score, per question, relative to the baseline score for that question. As an example, many participants

answered Question 8 differently after playing Obsolescence, and their new rated scores were an average of .65 points on the Likert-scale closer to the baseline score of 4. The associated LO for each question is represented by their color.

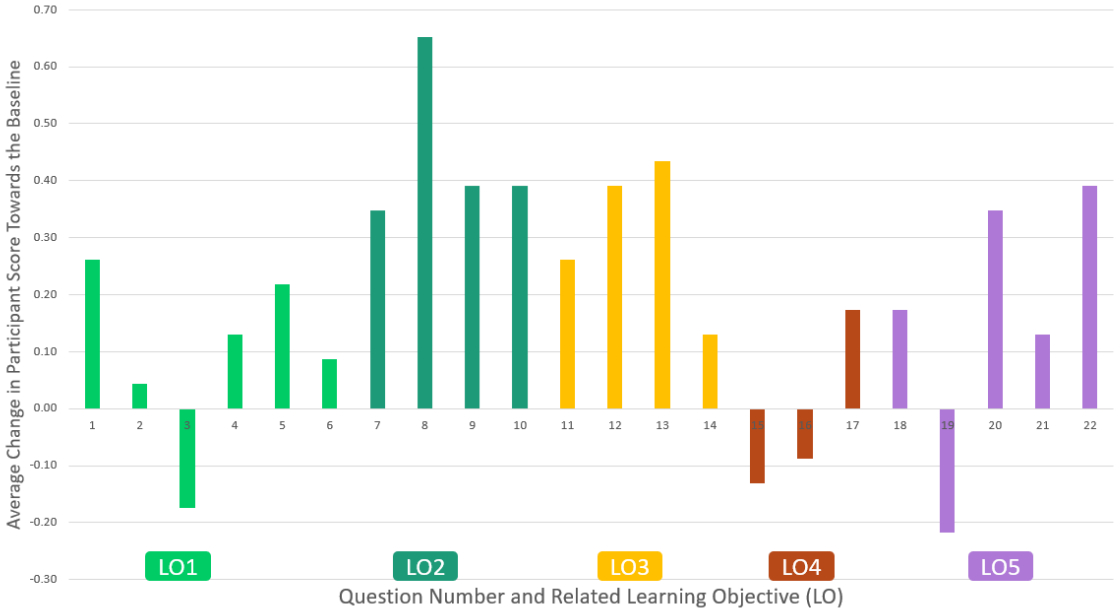


Figure 11: Average change in participant answers relative to the baseline, per question. Higher values indicate a stronger change towards the baseline.

On average, participants demonstrated a change in their answers in the post-survey of .20 per question. In other words, after playing Obsolescence participants would adjust each of their answers by an average of .2 higher or lower on the Likert scale relative to the established baseline. A complete table on measured participant learning is included in Appendix C.

LOs 2 and 3 had the highest measured learning. This may be for the following reasons: LO 2 was designed to be taught at a lower level of learning, *Understand*,

perhaps making it easier for participants to learn. LO3, *Support and value increases to military 'Complexity Carrying Capacity,'* despite scoring the lowest on the reported learning, may have high measured learning for a related reason. Players may not experience an increase in their Complexity Carrying Capacity during Obsolescence, which would explain the lower reported scores, but that does not mean they did not learn of the value of having an increased ability to process complex situations. If the game creates an environment where the ability to handle complexity is critical, players may realize the value of LO3 even without being able to experience solutions themselves. The complaints about game usability and the lack of LO3 Tech Cards that likely lead to the lower reported score may indicate that the game mechanics outside of the Tech Cards were reinforcing the concept.

LO4, which measures changes in participant values related to strategic plans and roadmaps for AI, demonstrated no measured learning across all participants. This may indicate that the game was not encouraging players to make complicated or multi-turn plans. In informal discussions with participants during the development and pilot tests of Obsolescence, several reported using simple heuristics or strategies instead of significant planning. The game was designed to only reward strategies involving significant strategy and planning, but participants may have gravitated towards alternate play styles that did not reinforce the concepts behind LO4. It is also possible that the questions used to assess LO4 were flawed in some way and failed to accurately measure the learning that was self-reported by participants. Only 3 questions measured the learning from LO4, whereas

other LOs had 4 or 5 each. Having more questions may have offset some of the low measurements for LO4, or revealed potential issues.

Comparison Between Reported and Measured Learning

The second research question involves a comparison between the measurement of learning and the reported learning. From this experiment, the measured learning is supported and validated by the self-reported questions. These numbers were calculated based on comparing all five post-survey Part One questions with all 22 pre- and-post-survey questions. The correlation coefficient, R, was calculated off of an array of reported learning scores and an array of measured learning. This value indicates the degree to which two dimensions are related, and ranges from -1.0 to +1.0, where in this experiment a higher positive correlation is desired. The t-score was generated using R and the total number of observations, and from those two values, a P-value was generated using a two-tailed t-distribution. P values under .05 indicate that the correlation between the two arrays is statistically significant.

On average, players reported spending less than 30 minutes in-game, with a standard deviation of 16 minutes. The players who reported high levels of learning, as determined by the top 40%, averaged 36 minutes in-game, played 3 rounds of Obsolescence, and had no significant difference from the rest of the population, in their answers to post-survey Part Two, Usability/Enjoyment.

Average measured learning vs average reported learning, per participant

This research found a +.58 correlation between the measured learning and the reported values for learning (P=.0041). When measuring any change in answers, not just changes towards the established baseline, the correlation is even stronger, at +.69 (P=.0001). Figure 12 shows a graphical representation of the correlation, using normalized values for the averages of each participant's reported learning and their measured learning. One participant did not answer the reported learning questions.

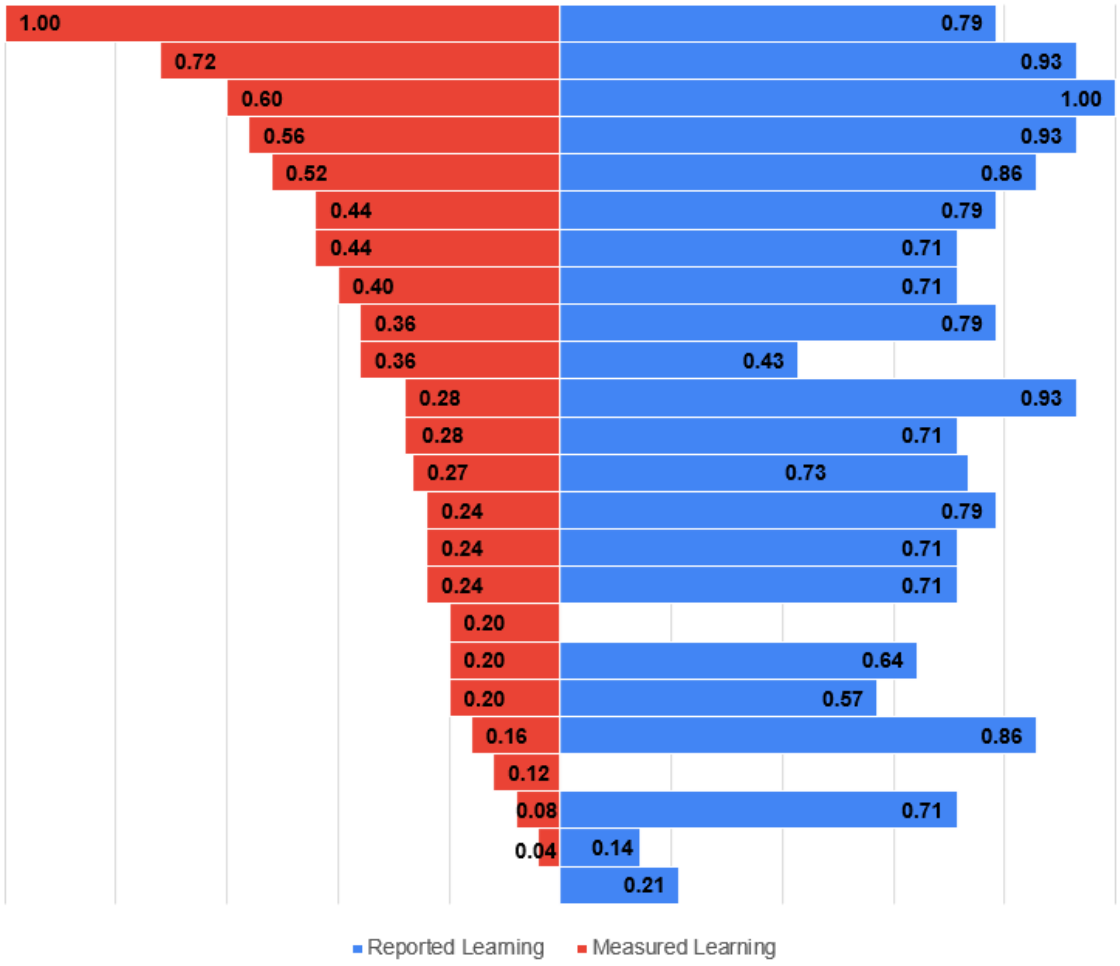


Figure 12: Graphical representation between measured and reported Learning, normalized, per participant

Average measured learning for a LO vs reported learning of the LO

Each of the 22 game-assessment questions in both the pre- and post-survey is associated with one of the 5 LOs. Table 8 displays the results when the same correlational analysis is conducted individually for each LO. From that analysis, only LOs 1 and 5 demonstrated strong significant correlations. For each question about LO1, participants

post-gameplay adjusted their answers an average of .11 points towards the baseline. This had a +.52 correlation with their reported level of learning for LO1 (P=0.0118). For the questions about LO5, post-gameplay scores were, on average, .20 points closer to the baseline. This measurement of LO5 had a +.46 correlation with participants' reported level of learning (P=.0276).

LO	1	2	3	4	5	Total
R	0.5186	0.17663	0.21747	0.42036	0.46659	**0.58204**
t	2.7125	0.80254	0.99641	2.07188	2.35922	**3.20103**
P value	0.01272	0.43082	0.32988	0.0502	0.02761	**0.00412**
Measured Learning	0.09	0.45	0.30	-0.01	0.17	**0.20**
std Dev	0.41	0.44	0.52	0.49	0.36	**0.27**

Table 8: Relations Between Reported and Measured Learning Per LO

Correlation results for LOs 2 and 3 had a P-value over .05 and are therefore not significant. LOs 2 and 3, however, displayed the highest average change in participants' answers. LOs 2 and 3 dealt with understanding the importance of AI technologies in the increasing complexity of warfare. Participants may not have felt any change in their opinions, viewpoints, or knowledge, but some change may have occurred. It is also possible that the concepts behind the LOs were taught but not the language. This would explain why participants did not feel like they understood the LOs as written, but did demonstrate an understanding when given more understandable scenarios.

Engagement

In section 2 of the post-survey, participants were asked to report on Obsolescence's enjoyability their perceived engagement. One question asked about how

much fun they had, one about the ease of access and usage, and one asked participants to estimate how much time they spent in-game. Other optional questions allowed participants to elaborate on why or what they felt the game did well or poorly concerning engagement and enjoyment.

Figure 13 and Figure 14 chart participant responses to questions involving their perception of usability and enjoyment. Most participants reported that the game and surveys were easy to access and use, but were neutral on the game's entertainment value. On average, participants reported a score of 3.47on a 1-5 scale for ease of access, and a 3.38 for entertainment.

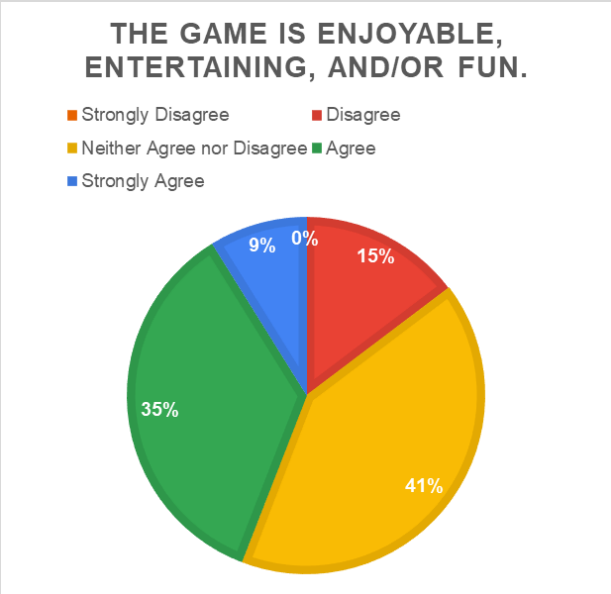


Figure 13: Responses to Post Survey Part Two-Question One

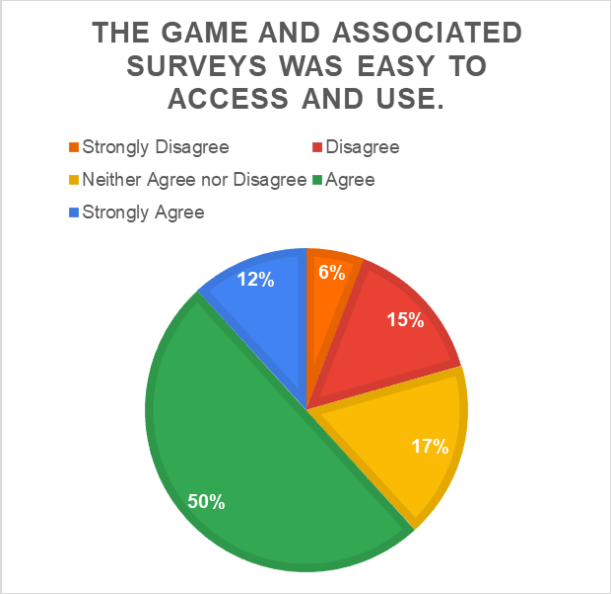


Figure 14: Responses to Post Survey Part Two-Question Two

Table 9 and Figure 15 display the answers to the entertainment and engagement questions alongside each participant's average reported learning score and the logged number of games they played. Higher reported enjoyment correlated positively with reported learning (r=+.55, P=.00052). This result reinforces the expectation of educational serious games. In addition, higher reported scores on ease of access and time in-game also correlated to reported learning. (r=+.50, P=.0018; r=+.48, P=.0029).

	Avg Reported Learning	Enjoyable/ entertaining /fun	Ease of Usage	Time in App	Number of Games Played
Correlation to Enjoyment	**0.554**	**1**	**0.52237**	**0.61399**	**0.21984**
t	3.82278	N/A	3.51909	4.46854	1.29457
P value	0.00052	N/A	0.00122	7.9E-05	0.20395
Correlation to Reported Learning	**1**	**0.554**	**0.50695**	**0.48656**	**0.15888**
t	N/A	3.82278	3.37853	3.19936	0.92442
P value	N/A	0.00052	0.0018	0.00292	0.3616

Table 9: Correlation between reported learning and Part Two of Post-Survey

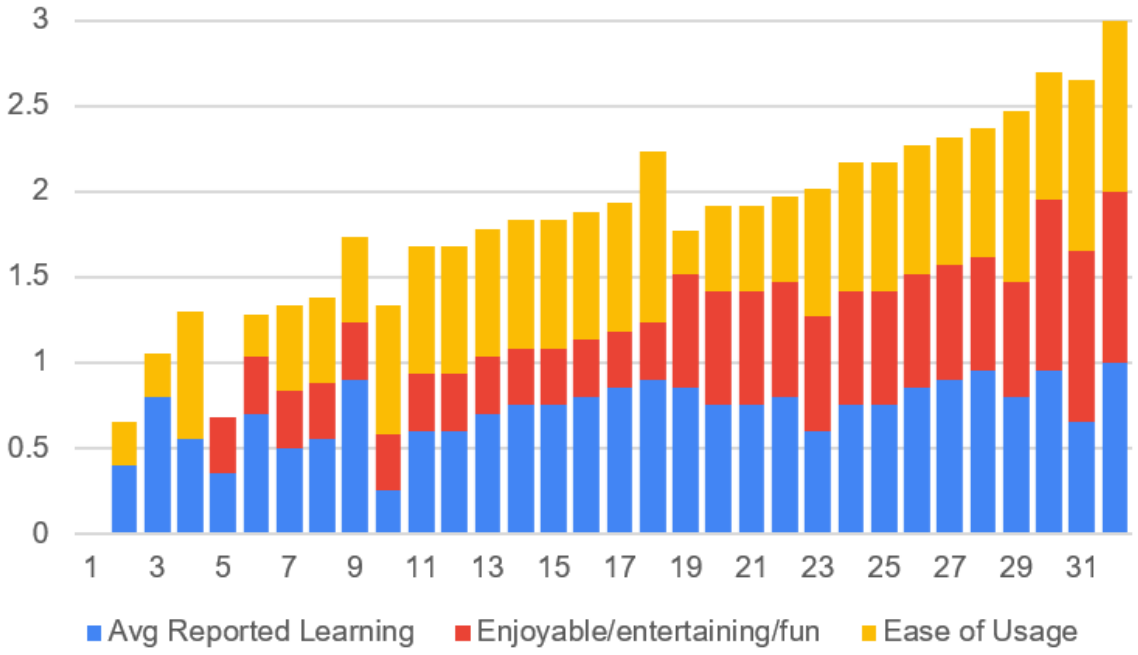


Figure 15: Comparison of reported learning and Part Two of Post-Survey. Higher values in one dimension correlate with higher values in other dimensions

While these results look like they indicate causality between enjoyable, easy-to-use games and higher levels of learning, this may not be the case. When compared

against the measured learning instead of the reported learning, the correlation drops significantly, and the P values rise significantly, as seen in Table 10. While the reported learning is correlated to the reported engagement, the correlation may not occur from higher engagement causing greater learning, but from a hidden factor.

	Enjoyable/ entertaining /fun	Ease of Usage	Time in App	Number of Games Played
Correlation to Measured Learning	**0.16898**	**0.26015**	**0.22339**	**0.09238**
t	0.7473	1.17438	0.99899	0.40442
P value	0.46317	0.25339	0.32917	0.68999

Table 10: Correlation between measured learning and Part Two of Post-Survey

The correlation may instead be explained by individuals displaying a natural bias towards higher or lower answers when asked to rate any experience, regardless of the specific question. This is an alternate explanation for the correlation found between all Reported Learning and Reported Engagement answers in the post-survey. As an example, a participant in a good mood might feel positive towards any question asking how they feel and any question asking if the time they just spent was well-spent. This may also have affected the questions in the Measured Learning questions of the post-survey, and therefore also the measurements of learning, but it appears to be less of a factor given the lower correlations to the measured learning.

Conclusion and Other Notes

The results from this experiment show clear correlations between a standard questionnaire format and the novel methodology intended to capture a more objective measurement of the game's educational value. Both measures reported that the game itself was successful, at least in part, at teaching its intended LOs. The game logs collected as part of the post-survey did not end up providing statistically significant data but can aid theories for improving both game and survey design and methodology.

For instance, from the game logs and the short responses, many participants either did not choose to take the tutorial or did not realize it existed. An analysis of the game logs shows that up to 14 players did not go through the game's tutorial. Recommendations for improvement such as "maybe include a tutorial" indicate some players did not notice the tutorial button. At least one player purposefully chose to skip the tutorial and "just wanted to play the game." Many of the players who did not take the tutorial deliberately lost their first game. That is, the game logs did not show them making any significant moves, instead just ending their turn and watching how the AI opponents played. This can be a viable strategy for learning a game and should be factored into game and experiment design. This may also have been motivated by participants seeking primarily to enjoy the game and not seeking to learn from the game.

Participants reported Obsolescence was a successful teaching tool, rating it a 3.8 out of a 5-point Likert scale. Individually, each LO was also deemed to be at least partially taught, with average reported learning ranging from 3.6 (LO3) to 4.0 (LO5).

Using the methodology developed to measure the learning from Obsolescence, participants demonstrated changes in their responses to real-world scenarios averaging .2 points closer to the baseline. Those two metrics are significantly and positively correlated across all participants at +.58 (P=.004). Lastly, participants were given unlimited time to play the game, resulting in average gameplay of 30±16 minutes.

The analysis here supports the idea that Serious Games are an effective teaching tool, and that Obsolescence, in particular, can teach players its LOs. This research did not examine if another medium may teach the same LOs to a greater degree, but instead used a baseline calibrated by SMEs to measure learning. Both the measured learning and the reported learning agreed that participants did experience learning, but the results from the self-reported learning did not differentiate much between LOs and may be more indicative of the participants' particular rating tendencies than an objective assessment. This is an issue with any self-reported survey, but the data from this experiment indicate that researchers may mitigate it by following the methodology in Chapter 4.

VI. Conclusion

Chapter Overview

This chapter summarizes the work conducted throughout this research including the design, study, and novel analytical approach of the educational SG Obsolescence. It reiterates a summary of the observations and conclusions found from this experiment, goes over the research contributions of this body of work, and discusses avenues for future work.

Research Summary

This research aims to determine the educational value of the SG Obsolescence and examine the value through both self-reported learning metrics and more direct measurements of learning. The overall goal of this research involved two parts. It produced a viable educational SG aimed at addressing DoD needs. The game was designed to meet LOs related to topics critical to the DoD's future success. Specifically, the game teaches teach values, concepts, and frameworks for decision-making in military domains when AI technologies are involved. AI technologies are rapidly becoming critical to warfighting capabilities, and there are many applications of such technologies that could cause significant disruption to the current military environment. However, in order to determine the educational value of the game, an assessment methodology was required.

Therefore, this research also succeeded in furthering the science behind SG evaluation. There are few measuring tools designed to objectively measure an SG on its own. There are also only limited tools to assist in measuring the learning of high-level concepts, a strength of SGs. Currently, most research relies on the opinion of professional educators or on self-reporting to evaluate SGs that teach high-level concepts. The methodology designed for this research measures learning without reliance on self-reporting or an instructor. Metrics from a self-reporting portion of the survey were used to allow a comparison between the novel assessment methodology and traditional SG assessment practices.

Obsolescence was found to teach its LOs without additional readings, instruction, or follow-up discussion groups. The methodology for objectively evaluating Obsolescence's effectiveness correlated with the reported measurement and supported the educational value of the game.

A total of 48 participants contributed data to this experiment. On average, they reported that the game taught its LOs, (3.8/5) was neither enjoyable nor disagreeable (3.4/5), and was moderately easy to access and use (3.5/5). Participants spent anywhere from 10 minutes to an hour in-game, and most completed the game and associated surveys in one seating. The players with the highest levels of learning spent an average of 36 minutes in-game, played three rounds of Obsolescence, and did not have a difference in enjoyment or engagement compared to the rest of the population.

The methodology for assessing learning using the delta between pre- and post-survey scores correlated with levels of reported learning (+.58). Participants, overall,

adjusted their responses in the post-survey closer to the correct answers by an average of .2 points per question (SD=.27).

Obsolescence taught some of its five LOs more than others. Post-survey questions related to LOs 2 and 3 had significant changes in participant answers, with participants changing their answers towards the correct responses at the average rate of .45 and .3 per question, respectively (SD=.44 and .52). The measurement of those two LOs did not correlate strongly or significantly to those LOs' reported learning. LOs 1 and 5, however, did have strong and significant correlations between the measured and reported levels of learning. (R=+.51, +.46, P=.012, .027). Those LOs dealt with valuing VVT&E and increased international monitoring.

Participants were asked to give short comments in the post-survey on the educational effectiveness of Obsolescence. The two most consistent types of comments were about increasing the playability of the game and incorporating more mediums into the educational experience. For instance, players thought that having a smoother UI or better tutorial would have both improved the fun of the game and the educational value. They also suggested tying in further readings or a breakdown of performance related to each of the LOs. If the game were to be used outside of an experimental setting, it would benefit from an attached workshop, course, or other material, as well as a graphical and user interface update.

Research Contributions

This research has made the following contributions to the field of educational games:

Scalable and Available Educational Tool:

Obsolescence, as it currently stands, has demonstrated the capability to teach its intended LOs. This experiment, by itself, did not determine the longevity of the learning nor the significance, only that some degree of learning was achieved. Obsolescence can be used as a 30-minute stand-alone experience to teach about strategic-level values and perspectives related to potentially disruptive AI technologies. This research makes no comparisons between the educational benefits of Obsolescence and other materials teaching the same concepts. The value of the game can likely be greatly enhanced when paired with other content or modes of learning. Informal feedback after the experience confirms the greater potential for Obsolescence when paired with other modes of learning. Two of the educational courses sourced for the study requested a follow-up discussion from the author about the topics covered in-game. Students in those classes confirmed that having an instructor cover the topics using the game as supporting material greatly enhanced the lessons taught by the game.

Methodology for Assessment and Evaluation of Educational Game Performance:

The game assessment methodology outlined in Chapter 4 of this paper provides an objective measurement of behavioral changes caused by gameplay. This methodology is designed to improve on the current state of SG assessment in the following ways:

1. The measurement does not rely on self-reporting from study participants.

   From the literature, self-reported learning is not reliable, and it also requires all participants to have the introspective skills to understand what learning did occur. The methodology includes statistics on self-reporting primarily to error-check the more objective measurements of learning.

2. It does not require a control group.

   Many game assessment studies have a low population size or are not able to create an equivalent environment for a control group. This methodology uses an established baseline to evaluate changes in participant behavior, values, and knowledge, instead of comparing results against a population that does not play the game.

3. All assessment can be done at scale and does not rely on the judgment of an individual.

   The study methodology does not necessitate that the researchers conduct interviews, record direct observations, grade survey answers, or participate in the

game. This not only standardizes the process but allows for the assessment to scale with any population size.

4. It may be applied to many games used for education.

    The methodology only requires that the game has associate LOs. It can apply to commercial games adapted for educational purposes or to SGs produced specifically to teach a lesson. While it can incorporate game logs, it does not rely on them. At this point, the methodology has not been tested on other games, but none of its characteristics are particular to Obsolescence.

5. It can evaluate learning on Bloom's affective taxonomy and high-levels of the cognitive taxonomy.

    Games that aim to teach affective concepts have few options for evaluating their success. This methodology gives a formal process for creating an experimental procedure and analysis plan that determines if a game has changed players' feelings, emotions, attitudes, or values.

Scalable Process for SG Hosting and Testing:

Experiments utilizing SGs are common at both AFRL and AFIT. Prior to this research, each SG would require its own infrastructure to support its hosting and distribution. Experiments were not always remotely accessible, either for researchers or participants. This research has worked closely with Hanger18 to produce a repeatable, scalable, and accessible process to host other SGs as a web app or server. Furthermore,

the authentication and data collection used for this experiment prompted the push towards a centralized and standardized cloud-based product to act as a wrapper on any further experiments conducted using the processes pioneered by this research.

Observations on Research Procedures and Lessons Learned

The process of designing, building and modifying Obsolescence had several setbacks and flaws that could have been avoided. First, the game suffered from over-ambitious game mechanics and complexity. This was in part due to the digital design. Creating the game for digital consumption makes it harder to change large aspects of the design. For initial development, Obsolescence could have utilized a tabletop mockup with physical tokens and a game master to simulate the game rules. This would have enabled the game rules to have been more fully visualized and fleshed out before any effort to put them down in code occurred.

In addition, most of the planned development for the game was assigned to mechanic creation and game functionality. In reality, approximately half of the development time was spent adjusting aspects of the game's visual and audio design, readability, UI layout, and other non-essential aspects of the game. For instance, having an intuitive method to read, select, and take actions on Tech Cards was vitally important to the pilot study participants, but also required more development hours than implementing most of the cards' game mechanics. Obsolescence's development needed both a greater focus on UI elements and a greater design budget for such features.

Some features of Obsolescence likely did not perform as designed. For instance, participants reported low levels of learning for LO3 and quoted the lack of helpful Tech Cards as the reason. There are Tech Cards designed to aid learning of that LO, however, given how late in the game they appear, players would rarely utilize those mechanics. While it is helpful to create a mapping of game mechanics to educational objectives, researchers need to ensure that the design and mechanics work as intended on a realistic player-base sample.

The software used to code Obsolescence, GameMaker Studio, prides itself on being easy to learn and allowing developers to make prototypes quickly [33]. While it did perform well for this experiment, the choice to use this particular software limits the future development and sustainability options for Obsolescence. Using a more popular engine, such as Unity, would have enabled other researchers and developers to build off of the game significantly easier.

The surveys used to collect participant data could also have been improved. The calibration from the SMEs ended up proving a significant difference from the initial intent of the questions. The intended scoring did not accurately represent how SMEs would value each question. This is likely due to the subjectivity introduced when translating documented recommendations into Likert-style survey questions. In addition, the language in these questions was often more technical than the language in the game. This may have led to errors in the game's assessment; future research should maintain the same level of language throughout the material and the assessment.

Lastly, the method for data collection was created and utilized mainly out of necessity. Future research will likely use a far more streamlined process. The research

could have benefited from direct control over the survey site and the game hosting site. Direct metrics on individuals who visited the game's website were not available, and the surveys relied on a jury-rigged system to track control numbers. Using a more formalized system would both allow more meta-data to be collected and reduce the number of un-linked surveys.

Future Work

This research could be improved and expanded upon in several notable ways. Listed below are three areas of interest concerning this work:

Test Obsolescence Using Other Educational Evaluation Methodologies:

Completing a full analysis of Obsolescence as an educational tool would likely include running similar experiments with different evaluation frameworks. This experiment was conducted entirely virtually and utilized no feedback from evaluators at the job sites or schoolhouses of participants. In the future, Obsolescence could be evaluated using subjective, yet powerful, measurements from SMEs, course instructors, or job evaluators. Those trained personnel could assess the amount of understanding of the LOs individuals demonstrate after playing the game. Comparing their conclusions with the findings from this research could greatly strengthen these experimental results. In addition, while Obsolescence has been used in two courses so far, neither course taught content similar to any of Obsolescence's LOs. The addition of Obsolescence was

not part of the coursework that the course instructor designed, and there are no current plans to use Obsolescence in teaching future sessions of those classes.

Obsolescence has, so far, been tested with two important caveats. First, it has not been used by participants who are expected to directly apply the skills they learned from the game. Using Obsolescence as training for specific jobs that would benefit from the game's educational goals would allow future researchers to assess the game based on objective measurements of job performance increases/decreases after playing the game.

Secondly, the game has always been tested without supporting materials or alternate modes of learning. Incorporating the game specifically and intentionally into a course about AI or military strategic studies might drastically improve the educational value of the game. This would have to be done by a researcher familiar with both Obsolescence and the target audience but would return data from a study more closely replicating a real-world usage of the SG.

Lastly, because this research has established a relatively objective baseline of game value, Obsolescence can be used as a known environment to test other assessment methodologies using the same subject recruitment plan and hosting techniques as this research.

Use the Same Educational Evaluation Methodology from this Study to Test Other Games:

The methodology developed and used to test Obsolescence is likely applicable to many other games. While in this case, it has shown to be useful for SG assessment,

running a similar experiment on other games would further specify the strengths and weaknesses of the assessment methodology. Researchers can follow the steps outlined in Chapter 4 to assess a different game, taking into consideration the new game's LOs. Following the same methodology would allow the two games to be measured with a standardized scale. If the games share one or many LOs, they could be directly compared.

Further Integration with Game Logs:

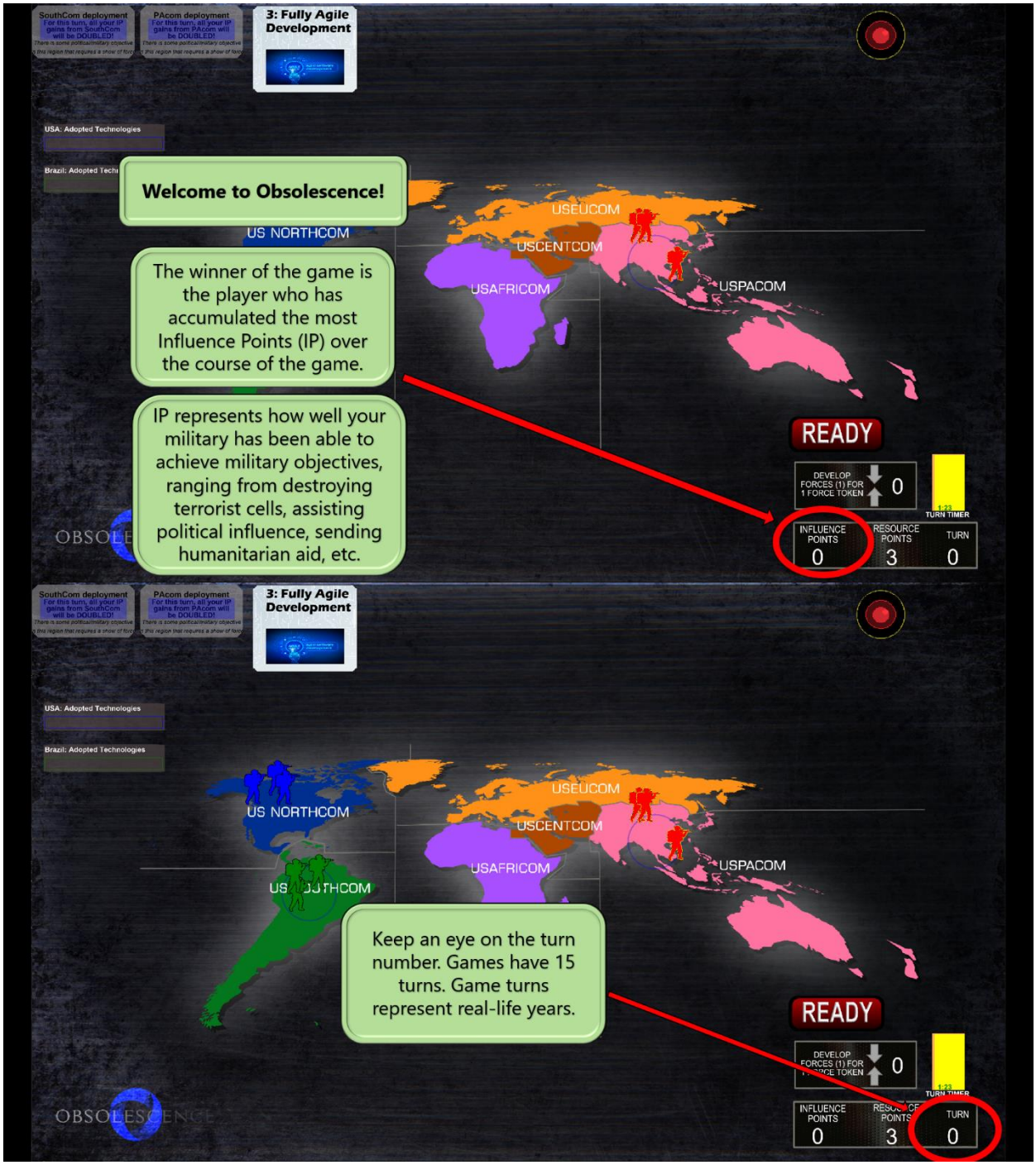
This experiment collected and utilized logs produced by Obsolescence. However, the log data was neither sufficient for serious statistical analysis, nor was the main focus of this research to look at in-game metrics. Future studies could focus on collecting and using game logs for several benefits. Using robust game logs of in-game actions could allow educators to move away from the survey questionnaire format altogether. If in-game behaviors could be tied closely to measurements of learning, player learning could be assessed based only on their in-game performance.

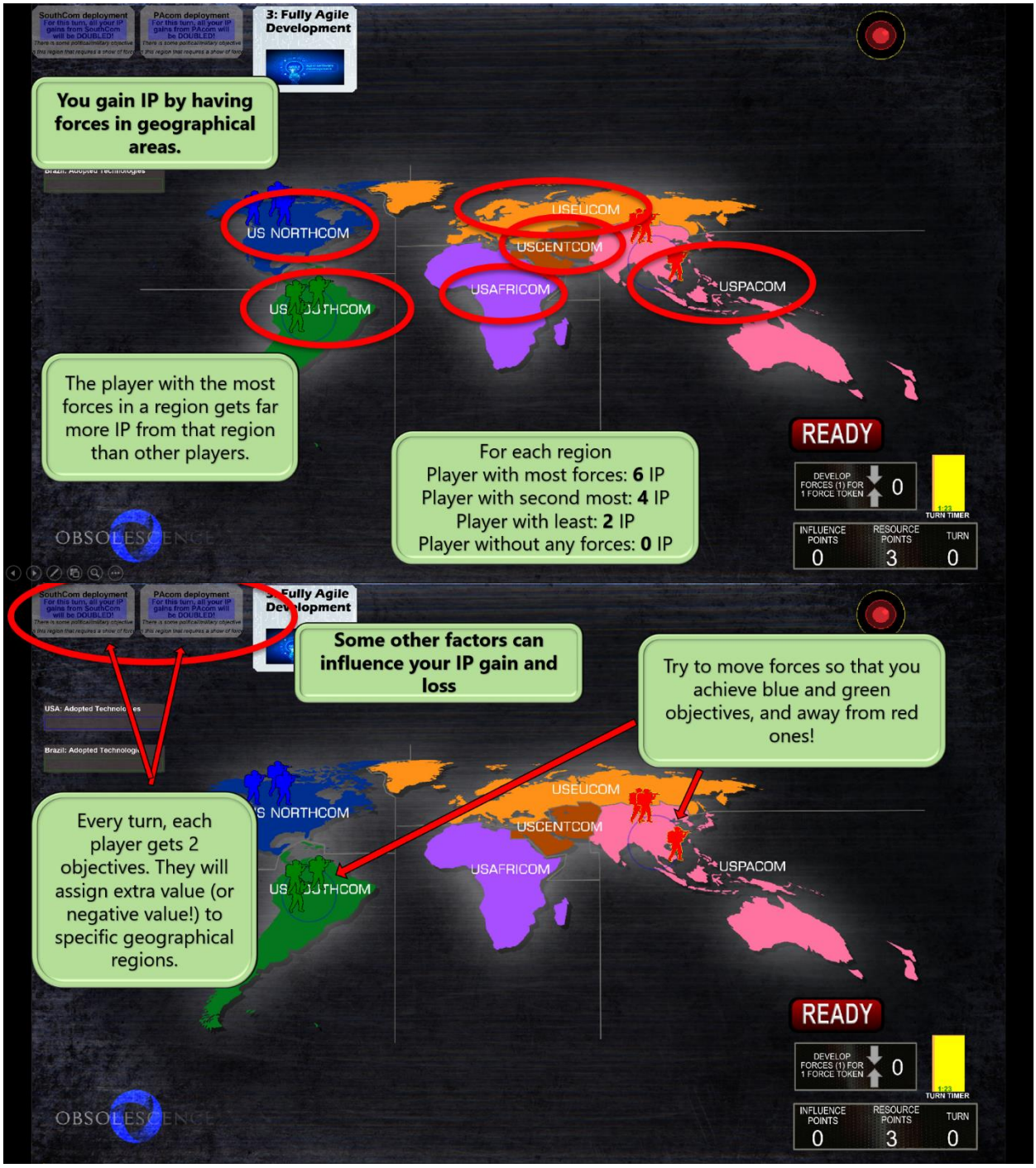
According to participant responses from this study, taking the surveys constituted roughly a third of their total time for the study. Further research focused on tying in-game logs might be able to make the game more accessible, and therefore a better tool for educators. Additionally, a larger data set would allow for significant modifications to Obsolescence to improve the educational value of the game. Data from this experiment was not detailed or significant enough to indicate which in-game actions correlated to players with higher levels of learning. An analysis of the game logs could aim at optimizing the game variables to encourage the best learning paths for players.

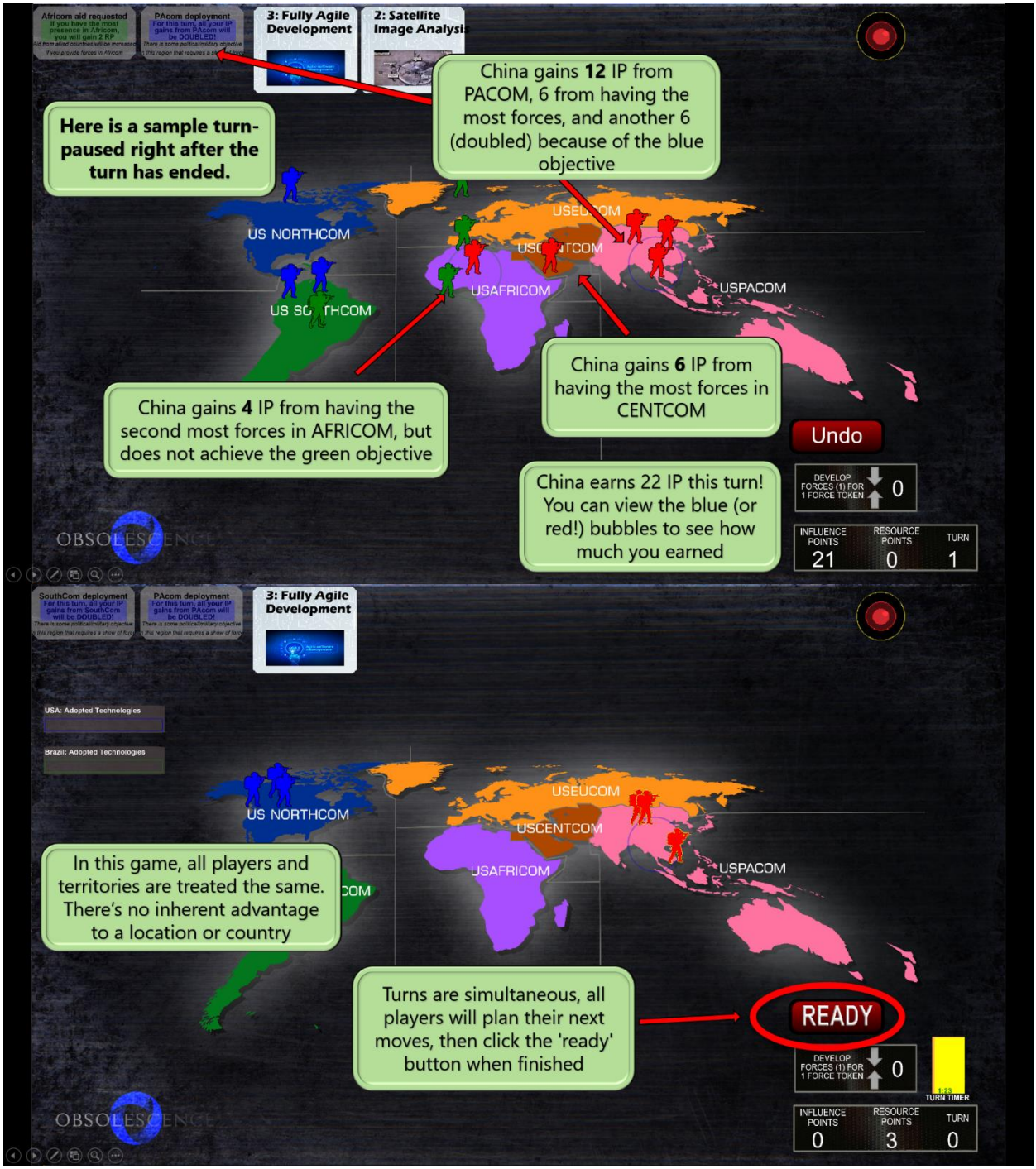
Final Thoughts

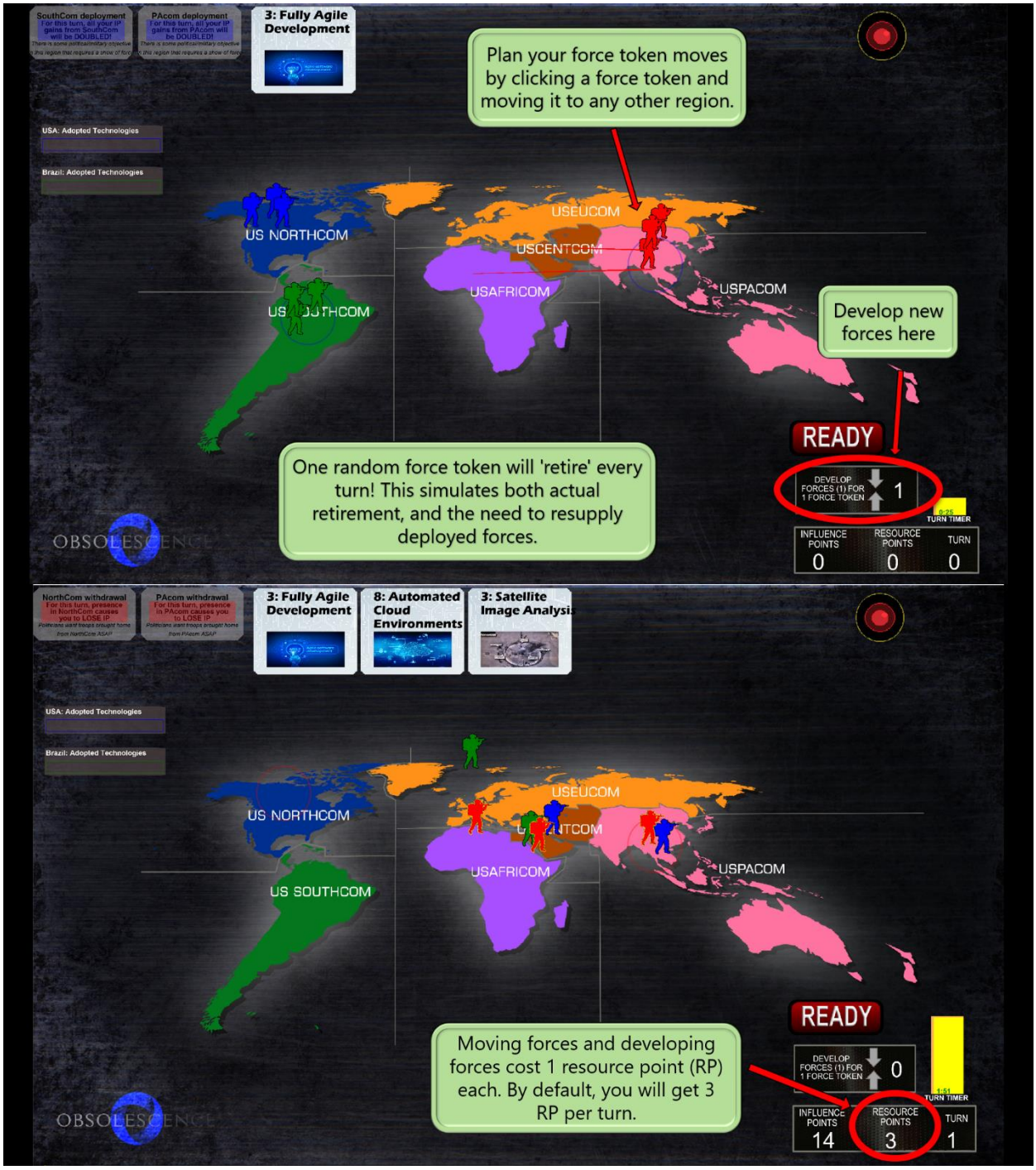
This research demonstrates the potential of Obsolescence as an educational tool, using both traditional measurements of learning and a novel game assessment methodology. The game was tested without any other medium of learning such as instructors, pre-reading, or discussion groups, and can teach value-based concepts about AI to an unfamiliar audience. The results show the potential for players, within 30 minutes, to shift their values and perspectives on real-world scenarios towards responses that authoritative sources and SMEs deem more correct. This result speaks toward the potential benefits all SGs may have for education and agrees with current literature supporting the potential of SGs. The methodology used to evaluate Obsolescence should be applied to a variety of other games for assessment and to further refine a standardized measurement of learning.

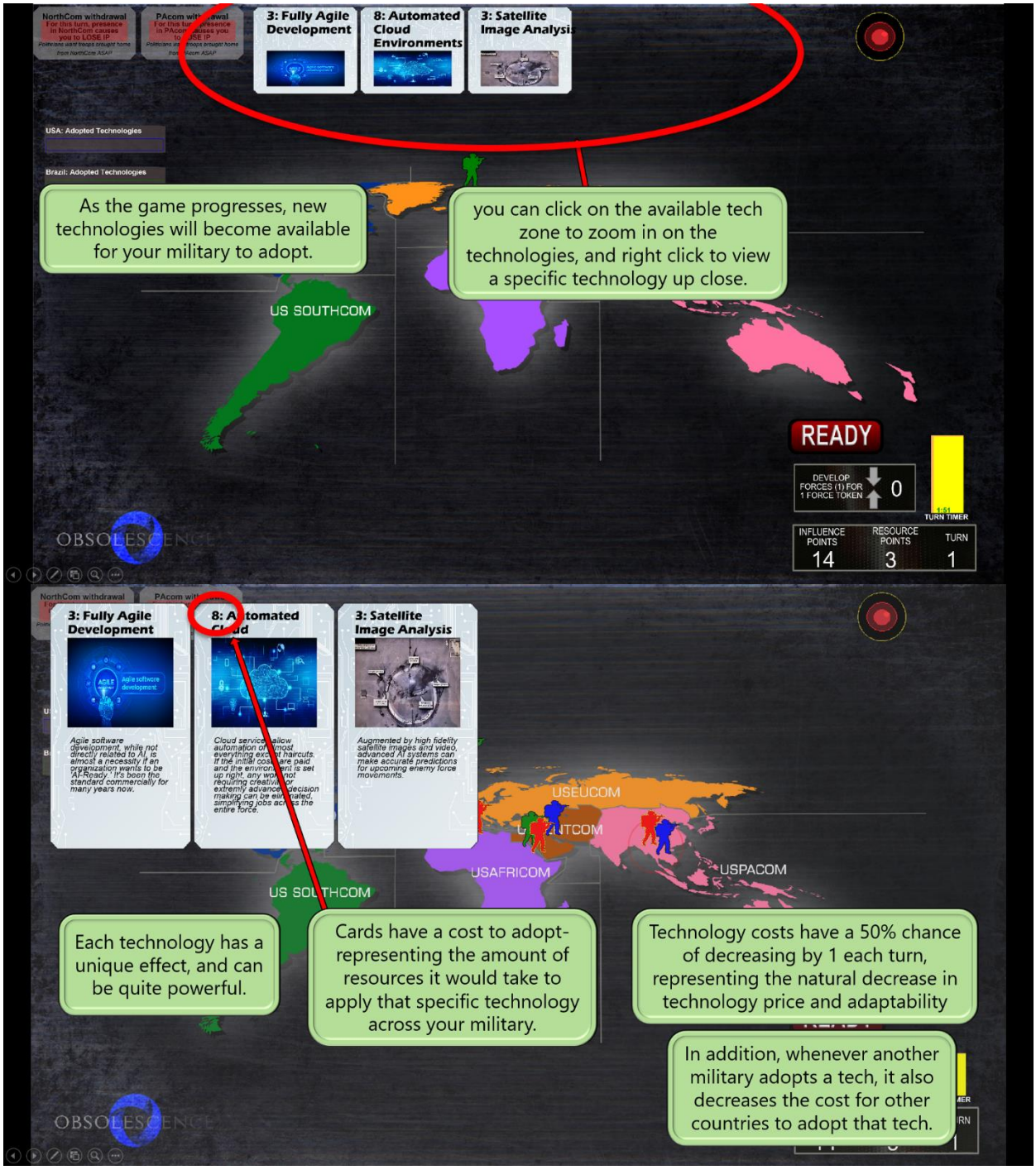
Appendix A: Obsolescence Rules

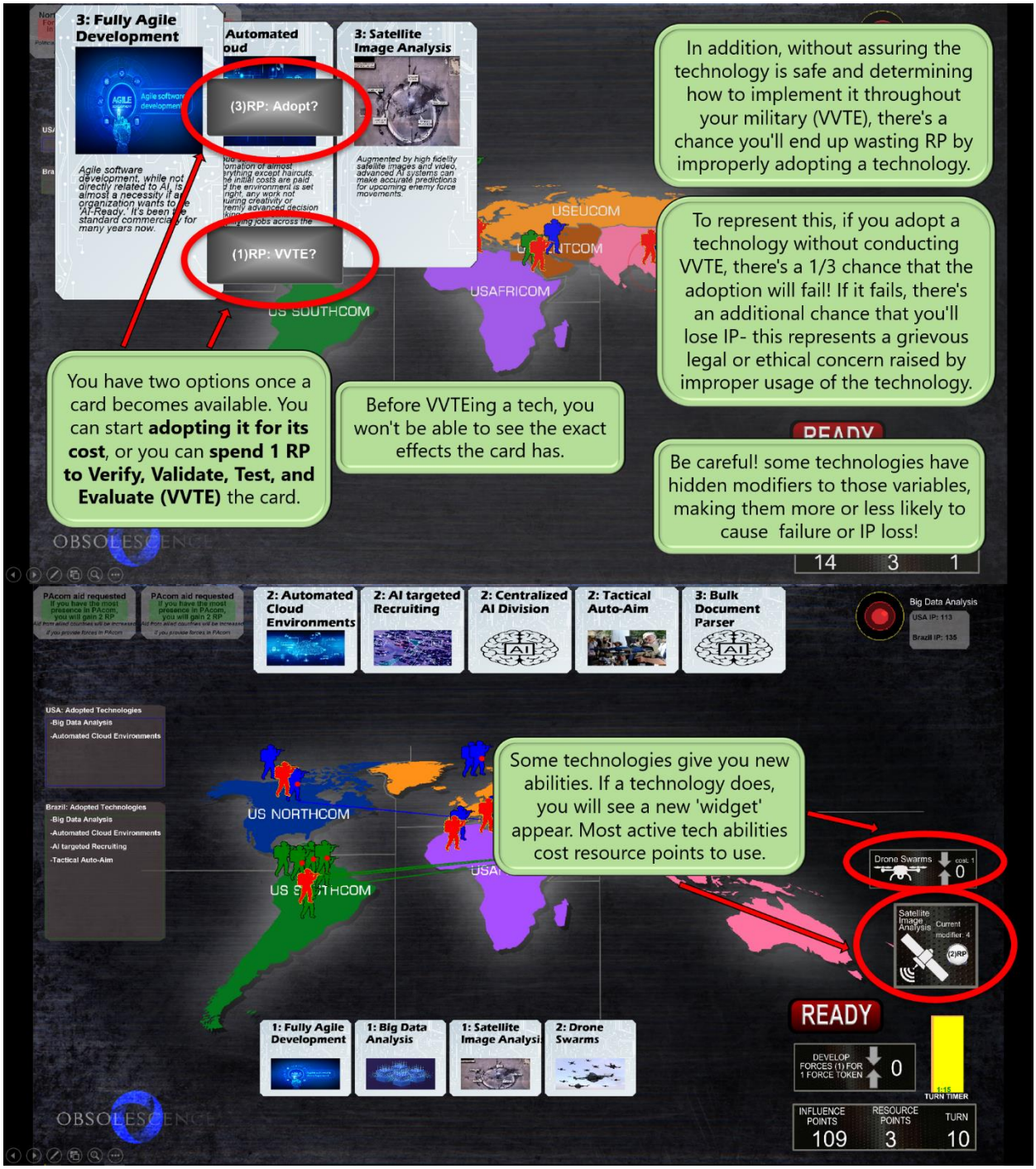


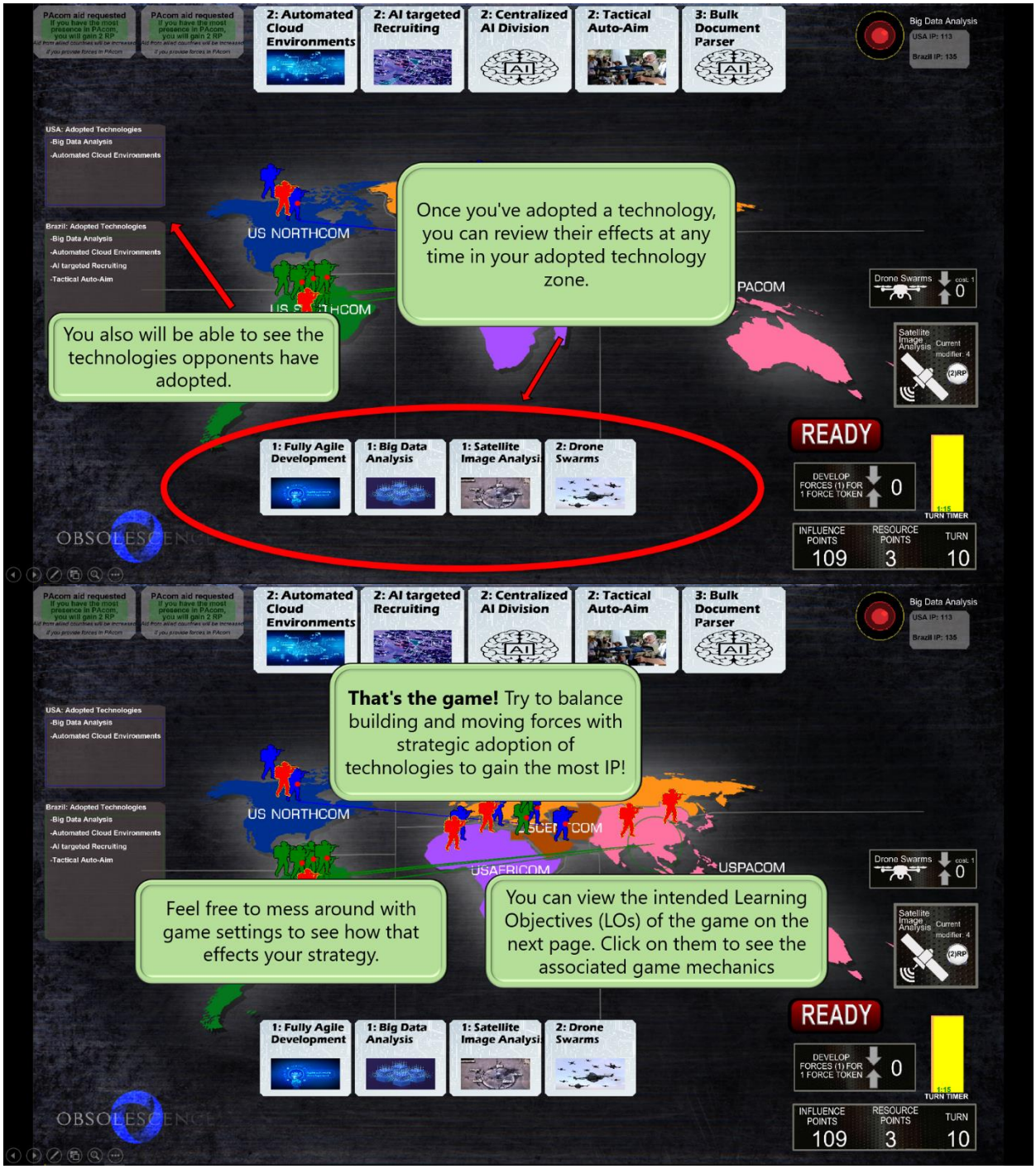

You gain IP by having forces in geographical areas.
US NORTHCOM
USEUCOM
USCENTCOM
USAFRICOM
USPACOM
The player with the most forces in a region gets far more IP from that region than other players.
For each region
Player with most forces: 6 IP
Player with second most: 4 IP
Player with least: 2 IP
Player without any forces: 0 IP
READY
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
0
TURN TIMER
INFLUENCE POINTS
RESOURCE POINTS
TURN
0
3
0
Some other factors can influence your IP gain and loss
Try to move forces so that you achieve blue and green objectives, and away from red ones!
Every turn, each player gets 2 objectives. They will assign extra value (or negative value!) to specific geographical regions.


Africom aid requested
If you have the most presence in Africom, you will gain 2 RP
PAcom deployment
For this turn, all your IP gains from PAcom will be DOUBLED!
3: Fully Agile Development
2: Satellite Image Analysis
Here is a sample turn- paused right after the turn has ended.
China gains 12 IP from PACOM, 6 from having the most forces, and another 6 (doubled) because of the blue objective
US NORTHCOM
USEUCOM
USCENTCOM
USAFRICOM
USPACOM
China gains 4 IP from having the second most forces in AFRICOM, but does not achieve the green objective
China gains 6 IP from having the most forces in CENTCOM
Undo
China earns 22 IP this turn! You can view the blue (or red!) bubbles to see how much you earned
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
0
INFLUENCE POINTS
21
RESOURCE POINTS
0
TURN
1
OBSOLESCENCE
SouthCom deployment
For this turn, all your IP gains from SouthCom will be DOUBLED!
PAcom deployment
For this turn, all your IP gains from PAcom will be DOUBLED!
3: Fully Agile Development
USA: Adopted Technologies
Brazil: Adopted Technologies
US NORTHCOM
USEUCOM
USCENTCOM
USAFRICOM
USPACOM
In this game, all players and territories are treated the same. There's no inherent advantage to a location or country
Turns are simultaneous, all players will plan their next moves, then click the 'ready' button when finished
READY
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
0
TURN TIMER
INFLUENCE POINTS
0
RESOURCE POINTS
3
TURN
0
OBSOLESCENCE


SouthCom deployment
For this turn, all your IP gains from SouthCom will be DOUBLED!
PAcom deployment
For this turn, all your IP gains from PAcom will be DOUBLED!
3: Fully Agile Development
Plan your force token moves by clicking a force token and moving it to any other region.
USA: Adopted Technologies
Brazil: Adopted Technologies
US NORTHCOM
US SOUTHCOM
USEUCOM
USCENTCOM
USAFRICOM
USPACOM
Develop new forces here
READY
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
1
TURN TIMER
INFLUENCE POINTS
RESOURCE POINTS
TURN
0
0
0
One random force token will 'retire' every turn! This simulates both actual retirement, and the need to resupply deployed forces.
OBSOLESCENCE
NorthCom withdrawal
For this turn, presence in NorthCom causes you to LOSE IP
PAcom withdrawal
For this turn, presence in PAcom causes you to LOSE IP
3: Fully Agile Development
8: Automated Cloud Environments
3: Satellite Image Analysis
USA: Adopted Technologies
Brazil: Adopted Technologies
US NORTHCOM
US SOUTHCOM
USEUCOM
USAFRICOM
USPACOM
READY
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
0
TURN TIMER
INFLUENCE POINTS
RESOURCE POINTS
TURN
14
3
1
Moving forces and developing forces cost 1 resource point (RP) each. By default, you will get 3 RP per turn.
OBSOLESCENCE


NorthCom withdrawal
3: Fully Agile Development
8: Automated Cloud Environments
3: Satellite Image Analysis
USA: Adopted Technologies
Brazil: Adopted Technologies
As the game progresses, new technologies will become available for your military to adopt.
you can click on the available tech zone to zoom in on the technologies, and right click to view a specific technology up close.
US SOUTHCOM
READY
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
0
TURN TIMER
INFLUENCE POINTS
14
RESOURCE POINTS
3
TURN
1
OBSOLESCENCE
3: Fully Agile Development
Agile software development, while not directly related to AI, is almost a necessity if an organization wants to be 'AI-Ready.' It's been the standard commercially for many years now.
3: Satellite Image Analysis
Augmented by high fidelity satellite images and video, advanced AI systems can make accurate predictions for upcoming enemy force movements.
USEUCOM
USAFRICOM
USPACOM
US SOUTHCOM
Each technology has a unique effect, and can be quite powerful.
Cards have a cost to adopt- representing the amount of resources it would take to apply that specific technology across your military.
Technology costs have a 50% chance of decreasing by 1 each turn, representing the natural decrease in technology price and adaptability
In addition, whenever another military adopts a tech, it also decreases the cost for other countries to adopt that tech.


3: Fully Agile Development
Agile software development, while not directly related to AI, is almost a necessity if an organization wants to be 'AI-Ready.' It's been the standard commercially for many years now.
Automated oud
3: Satellite Image Analysis
Augmented by high fidelity satellite images and video, advanced AI systems can make accurate predictions for upcoming enemy force movements.
(3)RP: Adopt?
(1)RP: VVTE?
In addition, without assuring the technology is safe and determining how to implement it throughout your military (VVTE), there's a chance you'll end up wasting RP by improperly adopting a technology.
To represent this, if you adopt a technology without conducting VVTE, there's a 1/3 chance that the adoption will fail! If it fails, there's an additional chance that you'll lose IP- this represents a grievous legal or ethical concern raised by improper usage of the technology.
You have two options once a card becomes available. You can start adopting it for its cost, or you can spend 1 RP to Verify, Validate, Test, and Evaluate (VVTE) the card.
Before VVTEing a tech, you won't be able to see the exact effects the card has.
Be careful! some technologies have hidden modifiers to those variables, making them more or less likely to cause failure or IP loss!
USEUCOM
USAFRICOM
US SOUTHCOM
14 3 1
PAcom aid requested
2: Automated Cloud Environments
2: AI targeted Recruiting
2: Centralized AI Division
2: Tactical Auto-Aim
3: Bulk Document Parser
Big Data Analysis
USA IP: 113
Brazil IP: 135
USA: Adopted Technologies
-Big Data Analysis
-Automated Cloud Environments
Brazil: Adopted Technologies
-Big Data Analysis
-Automated Cloud Environments
-AI targeted Recruiting
-Tactical Auto-Aim
US NORTHCOM
Some technologies give you new abilities. If a technology does, you will see a new 'widget' appear. Most active tech abilities cost resource points to use.
Drone Swarms
Satellite Image Analysis
Current modifier: 4
(2)RP
1: Fully Agile Development
1: Big Data Analysis
1: Satellite Image Analysis
2: Drone Swarms
READY
DEVELOP FORCES (1) FOR 1 FORCE TOKEN
0
TURN TIMER
INFLUENCE POINTS 109
RESOURCE POINTS 3
TURN 10



Additional game details:

- Tech Cards are divided up into three groups: Immediate, Near, and Future. Each turn, 0-1 cards are added to the pool of cards that can be adopted. For the first 5 turns, cards are only drawn from the Immediate group. Turns 6-10 draw from

both/either of the Immediate and Near groups, and the remaining turns draw from all three card groups.

- Players start the game with a set timer to their turn length, by default 1:30. Each game turn decreases the amount of time players have to take actions by 4 seconds.
- Upon clicking the menu button to start a singleplayer game, players are presented with links to the pre- and post-surveys. Those links are generated by the game using the Google Forms pre-filled URL, concatenating the players' control number, and game data for the post-survey, to the section of the URL that pre-fills out an answer.
    - https://docs.google.com/forms/d/e/1FAIpQLScYJTer7UKZujJjieZ-NADkx2ASgTtgAXhOvR9KfLWzEm4www/viewform?usp=pp_url&entry.826332445= += _data
- Game settings available for players to change:
    - Number of players
    - Income per turn
    - Initial Force count
    - Turn Timer
    - Random Seed
    - Max Techs per turn

Appendix B: Survey Questions

All questions only allowed responses on a 5 point Likert scale unless specified otherwise.

The scale ranged from (1) *strongly disagree* to (5) *strongly agree*.

Pre-survey and post-survey Section 3:

1. DoD technology advisors should ONLY be allowed to discuss technologies if they can fully explain the uncertain timelines and effects
2. DoD should stand up an annual, joint, multi-million dollar exercise to wargame AI technologies and risks from human-machine interactions
3. DoD should pull researchers from military projects to instead focus on developing more rigorous methodologies for testing AI systems
4. DoD budgets for AI technology should allocate about 40% of the total funds to be used for VVTE efforts
5. DoD should create a specific educational concentration for 'AI test and Evaluation Engineer,' in addition to teaching software VVTE in other concentrations.
6. DoD should not use, and should stop using, AI technologies that have not gone through a rigorous VVTE process
7. DoD should push culture changes, similar to the recent anti-extremism push, to encourage new technology usage
8. DoD should cease development and funding for military systems that are ill-equipped for AI-enabled warfare
9. Every CoCom staff should add a new member (similar to the staff JAG) that is an expert exclusively on AI systems
10. DoD needs to integrate AI systems, applications, and scenarios into all major exercises.
11. DoD should prioritize purchasing and utilizing technologies that are shown to improve digestion and understanding of large amounts of information
12. The responsibilities of Senior Executive and General Officers should increase to include: 1) Be inspired by AI and able to inspire the organization, and 2) Build and maintain an AI vision and strategic plan
13. In addition to all other anti-disinformation efforts, the DoD needs to stand up a 24/7 task force to combat AI-produced disinformation
14. DoD field commanders should use AI software for real-time decision support and Course of Action development.

15. DoD needs a 5 year strategic road-map for AI in order to not fall behind adversaries technologically
16. DoD needs to have EACH service organization stand up a team responsible for developing and maintaining a 5 year roadmap of AI warfighting technologies.
17. DoD should implement bi-annual portfolio reviews of ALL DoD investments in AI.
18. The Joint Chiefs of Staff should be briefed quarterly on developments in military and commercial AI technology
19. Defense Intelligence Agency (DIA) should monitor the state of all AI developments in other countries, even those developments where no military application is observed.
20. DoD needs to develop and follow a well-defined cohesive plan for international monitoring and engagement of AI technologies, even at the cost of increased international tensions
21. DoD should enforce, internationally and domestically, U.S. policy that only human beings can authorize employment of nuclear weapons.
22. DoD should develop and enforce international standards for VVTE of AI systems, even at the cost of increased international tensions

Post-survey Section 1:

23. Playing this game has aided your understanding of LO1: Almost all AI technologies need extensive and rigorous Verification, Validation, Testing, and Evaluation (VVTE) in order to be effective and/or low-risk
24. Playing this game has aided your understanding of LO2: The complexity of the world is increasing, in large part because of the effects of AI technology. Military decision makers will have more and more options, and more data with which to make decisions, likely degrading decision making
25. Playing this game has aided your understanding of LO3: Just as the complexity of military systems is increasing, the ability to understand and deal with vast amounts of information (Complexity Carrying Capacity) can be increased through utilization of specific AI technologies
26. Playing this game has aided your understanding of LO4: AI technology will impact all aspect of military operations- Every part of the DoD can benefit from a 5 year strategic plan/roadmap for emerging AI technology
27. Playing this game has aided your understanding of LO5: AI research and military adoption needs to be monitored in other countries, or else the DoD will quickly lose our technological advantage

Post-Survey Section 2:

28. The game is enjoyable, entertaining, and/or fun.
29. The game and associated surveys was easy to access and use.
30. If you had access or frustrations completing the game or surveys, please make a short note of the problems you encountered (Short Answer)
31. Approximately how much time did you spend playing the game? (Multiple Choice)
32. Do you have any recommendations to improve game enjoyability? (Short Answer)
33. Do you have any recommendations to improve the game's educational value, specifically of the above LOs? (Short Answer)

Appendix C: Additional Data Tables and Figures

Control Number	Avg Reported Learning	Enjoyable/ entertaining /fun	Ease of Usage	Time in App	Number of Games Played
799161		2	2	20	
1166668	4.8	5	4	45	3
172576	2.4	3	1	10	3
150544	5.0				
122440	4.4	3	4	30	6
1128128	4.2	3	4	20	1
1160	4.0	4	4	20	1
12880	3.6	5	5	60	4
1.302E+10	4.4	4	4	30	2
4160	2.6	2	2	20	1
119872	1.0	2	1	20	1
154600	5.0	5	5	60	
117625	4.0	4	3	30	4
132340	4.8	4	4	60	1
148300	3.4	4	4	10	4
10	4.2	4	5	60	9
129184		4	4	45	1
1200340	3.8	3	4	10	1
165520	2.0	3	4	10	1
192988	3.0	3	3	10	
143200	3.8	3	2	20	1
1144495	3.2	3	3	30	2
136960	4.0	3	4	20	1
122325	4.6	4	4	45	3
1142324	3.4	3	4	30	7
1216580	3.2	2	4	10	1
1110976	4.4	4	2	30	3
176545	4.6	3	3	30	3
1153216	4.2	4	3	30	1
161874	3.4	3	4	20	2
1199892	4.0	3	4	45	4
136961	4.0	4	3	10	1
18316	4.2	2	2	20	2
195040	4.6	3	5	10	1
1168300	4.0	4	4	30	1
Average	3.8	3.4	3.5	27.9	2.5

Table 11: All post-survey section 2 results and number of games played.

Question #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
10	0	0	0	1	0	0	0	1	0	0	0	-1	0	0	1	0	0	0	0	0	0	0
1160	1	0	-3	0	0	0	1	2	1	0	0	0	0	0	0	1	0	0	-1	0	0	0
4160	-1	-1	0	-1	0	0	1	0	0	0	0	1	0	1	-3	-1	0	0	0	0	0	0
12880	-1	0	0	1	0	-1	0	1	1	-1	0	0	1	1	0	0	0	-1	0	0	0	0
18316	0	0	0	0	0	0	0	2	1	1	0	1	0	0	0	0	0	0	0	0	0	0
117625	0	0	0	0	0	0	0	0	1	4	0	0	0	0	0	-1	-1	4	0	0	0	0
122325	2	1	0	1	0	0	1	2	1	1	0	1	0	0	0	1	1	0	0	0	1	1
122440	2	0	1	0	0	0	0	1	1	1	-1	0	1	0	0	1	0	0	0	1	1	0
129184	2	0	-2	0	1	0	1	1	0	0	1	0	2	0	-1	-1	0	-1	0	0	-1	0
136960	0	-1	-1	-1	0	-1	0	-1	0	0	0	0	0	1	0	0	1	1	0	0	0	0
136961	0	1	0	-1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	1
165520	-1	1	-1	0	1	0	2	0	0	0	0	-1	-1	0	1	1	-1	-1	0	0	0	-1
172576	0	-1	0	0	-1	0	-1	2	0	-1	0	1	4	-3	-2	-2	2	1	-4	0	2	0
176545	-1	1	0	1	1	0	1	0	1	1	0	2	0	1	0	0	0	0	0	1	0	1
195040	1	0	0	0	1	1	0	0	0	0	0	1	0	-1	0	0	0	0	0	0	0	0
799161	0	-1	-1	0	-1	0	0	-1	1	2	0	-1	0	0	0	0	0	0	0	1	0	2
1110976	0	0	1	1	1	-2	0	0	-1	0	1	1	-1	0	0	-2	0	0	0	0	0	1
1128128	0	0	2	0	0	3	0	3	0	0	5	1	1	1	0	0	1	1	1	1	0	1
1153216	0	1	0	0	1	0	1	0	0	1	1	1	1	0	0	1	0	1	-1	0	-1	0
1166668	2	1	0	0	1	0	0	0	0	0	0	1	0	1	1	0	1	0	0	2	0	1
1199892	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0
1200340	0	-1	0	0	0	0	0	0	0	0	0	-1	0	0	0	0	0	0	0	1	1	1
1216580	0	0	0	1	0	2	0	1	0	0	-1	1	0	0	0	0	0	-1	0	1	0	1
Total Improvement per question	6	1	-4	3	5	2	8	15	9	9	6	9	10	3	-3	-2	4	4	-5	8	3	9
Avg Improvement per question	0.26	0.04	-0.17	0.13	0.22	0.09	0.35	0.65	0.39	0.39	0.26	0.39	0.43	0.13	-0.13	-0.09	0.17	0.17	-0.22	0.35	0.13	0.39
Related LO	1						2				3				4			5				

Table 12: Complete data on participant changes towards the baseline. Measured per participant, per question, with aggregate scores. Participants who gave answers closer to the baseline after playing Obsolescence have positive scores depending on the degree

VII. Bibliography

[1] S. of D. D. M. T. Esper, “Secretary of Defense Remarks for DOD Artificial Intelligence Symposium and Exposition,” 2020.

[2] O. of the SecDef, “China Military Power Report 2019,” 2019.

[3] A. Bundy, “Preparing for the future of Artificial Intelligence,” *Exec. Off. Pres. Natl. Sci. Technol. Counc.*, vol. 32, no. 2, pp. 285–287, 2016, doi: 10.1007/s00146-016-0685-0.

[4] A. Solow-Niederman, “The Department of Defense Posture for Artificial Intelligence,” *Danielle C. Tarraf, William Shelton, Edward Park. Brien Alkire, Diana Gehlhaus, Justin Grana, Alexis Levedahl, Jasmin Leveille, Jared Mondschein, James Ryseff, al.*, 2021.

[5] R. T. Sataloff, M. M. Johns, and K. M. Kost, “DoD AI training and education strategy and infographic,” *JAIC*, [Online]. Available: https://www.ai.mil/docs/2020_DoD_AI_Training_and_Education_Strategy_and_Infographic_10_27_20.pdf.

[6] C. Fadel, “Multimodal Learning Through Media: What the Research Says,” 2008.

[7] R. Moreno and R. Mayer, “Interactive multimodal learning environments: Special issue on interactive learning environments: Contemporary issues and trends,” *Educ. Psychol. Rev.*, vol. 19, no. 3, pp. 309–326, 2007, doi: 10.1007/s10648-007-9047-2.

[8] H. F. O’Neil, R. Wainess, and E. L. Baker, “Classification of learning outcomes: Evidence from the computer games literature,” *Curric. J.*, vol. 16, no. 4, pp. 455–474, 2005, doi: 10.1080/09585170500384529.

[9] Edutopia, *James Paul Gee on Learning with Video Games*. 2012.

[10] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, "Taxomony of Educational Objective," *Taxon. Educ. Object.*, pp. 62–197, 1956.

[11] M. E. Linick and J. Y, "Hegemony: a Game of Strategic Choices, Rulebook." RAND Corporation., Santa Monica, Calif.:, 2020.

[12] U. Ritterfeld, M. Cody, and P. Vorderer, "Serious Games: Mechanisms and Effects.," *Routledge, London*, p. 6, 2009.

[13] R. E. Clark, "Learning from Serious Games," *Educational Technology*, vol. May-June. pp. 56–59, 2007.

[14] E. Schmidt *et al.*, "NSCAI Final Report." National Security Commission on Artificial Intelligence, 2021.

[15] S. Barab, M. Gresalfi, and A. Arici, "Why educators should care about Games," *Educ. Leadersh. J. Dep. Superv. Curric. Dev.*, vol. 67, pp. 76–80, 2009.

[16] P. Perla, *The Art of Wargaming: A Guide for Professionals and Hobbyists*. 1990.

[17] R. Blunt, "Does Game-Based Learning Work ? Results from Three Recent Studies Does Game-Based Learning Work ? Results from Three Recent Studies," *Training*, pp. 1–11, 2009.

[18] G. Gris and C. Bengtson, "Assessment Measures in Game-based Learning Research," *Int. J. Serious Games*, vol. 8, no. 1, pp. 3–26, Mar. 2021, doi: 10.17083/ijsg.v8i1.383.

[19] T. Sitzmann, "A meta-analytic examination of the instructional effectiveness of computer-

based simulation games," *Pers. Psychol.*, vol. 64, no. 2, pp. 489–528, 2011, doi: 10.1111/j.1744-6570.2011.01190.x.

[20] L. A. Annetta, J. Minogue, S. Y. Holmes, and M. T. Cheng, "Investigating the impact of video games on high school students' engagement and learning about genetics," *Comput. Educ.*, vol. 53, no. 1, pp. 74–85, 2009, doi: 10.1016/j.compedu.2008.12.020.

[21] L. W. Anderson *et al., A Taxonomy for Learning, Teaching, and Assessing: A revision of Bloom's Taxonomy of Educational Objectives*. New York, NY, USA: Pearson, Allyn & Bacon, 2001.

[22] O. J. Petit dit Dariel, T. Raby, F. Ravaut, and M. Rothan-Tondeur, "Developing the Serious Games potential in nursing education," *Nurse Educ. Today*, vol. 33, no. 12, pp. 1569–1575, 2013, doi: 10.1016/j.nedt.2012.12.014.

[23] A. Calderón and M. Ruiz, "A systematic literature review on serious games evaluation: An application to software project management," *Comput. Educ.*, vol. 87, pp. 396–422, 2015, doi: 10.1016/j.compedu.2015.07.011.

[24] M. Prensky, "True Believers: Digital Game-Based Learning in The Military," *Digit. Game-based Learn.*, pp. 1–18, 2001.

[25] L. A. Freeman, "Instructor time requirements to develop and teach online courses.," *Online J. Distance Learn. Adm.*, vol. 18, no. 1, pp. 1–15, 2015, [Online]. Available: https://www.westga.edu/~distance/ojdla/spring181/freeman181.html.

[26] "Interview with AFIT Acquisition Education Research Analyst." 2021.

[27] R. A. Defelice, "How Long Does It Take to Develop Training? New Question, New

Answers," *Associatino for Talent Development*, 2021. https://www.td.org/insights/how-long-does-it-take-to-develop-training-new-question-new-answers.

[28] C. Girard, J. Ecalle, and A. Magnan, "Serious games as new educational tools: How effective are they? A meta-analysis of recent studies," *J. Comput. Assist. Learn.*, vol. 29, no. 3, pp. 207–219, 2013, doi: 10.1111/j.1365-2729.2012.00489.x.

[29] K. Schrier *et al.*, "Assessment in and of Serious Games: An Overview," *J. Prof. Nurs.*, vol. 17, no. 2011, pp. 31–35, 2017, [Online]. Available: http://dx.doi.org/10.1016/j.sbspro.2011.12.025%0Ahttp://delivery.acm.org/10.1145/2490000/2484486/p1-bellotti.pdf?ip=218.104.71.166&id=2484486&acc=NO RULES&key=BF85BBA5741FDC6E.A4F9C023AC60E700.4D4702B0C3E38B35.4D4702B0C3E38B35&__acm__=1559114231_4422d168b.

[30] A. De Gloria, F. Bellotti, and R. Berta, "Serious Games for education and training," *Int. J. Serious Games*, vol. 1, no. 1, 2014, doi: 10.17083/ijsg.v1i1.11.

[31] M. Konaev, C. Husanjot, R. Fedasiuk, T. Huang, and I. Rahkovsky, "U.S. Military Investments in Autonomy and AI," no. October, 2020, [Online]. Available: https://cset.georgetown.edu/publication/u-s-military-investments-in-autonomy-and-ai-a-strategic-assessment/.

[32] G. Jensen and M. Largent, "Design for Maritime Singularity," *Off. Nav. Res.*, no. September, p. 31, 2017.

[33] Yoyogames, "Gamemaker Studio 2," 2021. https://www.yoyogames.com/en/gamemaker.

[34] R. Lathrop, "Introduction to AI & Intelligent Agents." 2020, [Online]. Available: https://www.ics.uci.edu/~rickl/courses/cs-171/cs171-lecture-slides/2016_WQ_CS171/cs-

171-01-Agents.pdf.

[35] B. W. Everstine, “Artificial Intelligence Easily Beats Human Fighter Pilot in DARPA Trial,” *Air Force Mag.*, 2020, [Online]. Available: https://www.airforcemag.com/artificial-intelligence-easily-beats-human-fighter-pilot-in-darpa-trial/.

[36] P. Tucker, “The Air Force Used AI to Operate the Radar on a U-2 Spy Plane,” *Def. One*, 2020, [Online]. Available: https://www.defenseone.com/technology/2020/12/air-force-used-ai-operate-radar-u-2-spy-plane/170813/.

[37] P. Singer, “Tactical Generals: Leaders, Technology, and the Perils,” *Brookings*, 2009, [Online]. Available: https://www.brookings.edu/articles/tactical-generals-leaders-technology-and-the-perils/.

[38] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science (80-.).*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.

[39] T. Ord, *The Precipice*. 2020.

[40] S. Russell, *Human Compatible*. 2019.

[41] B. Christian, *The Alignment Problem*. 2020.

[42] M. Skilton and F. Hovsepioan, *The 4th Industrial Revolution*. 2018.

[43] N. M. Chaillan, “It is time to say Goodbye!,” 2021. https://www.linkedin.com/pulse/time-say-goodbye-nicolas-m-chaillan/.

[44] A. Milani and V. Poggioni, “PLANNING IN REACTIVE ENVIRONMENTS,” *Comput. Intell.*, vol. 23, no. 4, pp. 439–463, 20007, doi: 10.1111/j.1467-8640.2007.00315.x.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
21-03-2022	Master's Thesis	Sept 2020 – March 2022

TITLE AND SUBTITLE
'OBSOLESCENCE:' EVALUATING AN EDUCATIONAL SERIOUS GAME ON ARTIFICIAL INTELLIGENCE IMPACTS TO MILITARY STRATEGIC GOALS

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Timothy C. Kokotajlo., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENG)
2950 Hobson Way, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENG-MS-22-M-039

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Intentionally left blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
Artificial Intelligence (AI) threatens to bring significant disruption to all aspects of military operations. This research develops a Serious Game (SG) and assessment methodology to provide education on the mindsets required for engaging with disruptive AI technologies. The game, Obsolescence, teaches strategic-level concepts recommended to the Department of Defense (DoD) from a compilation of reports on the current and future state of AI and warfighting. The methodology for assessing the educational value of Obsolescence addresses common challenges such as subjective reporting, control groups, population sizes, and measuring abstract or high levels of learning. The game's proposed educational value is tested using a pre- and post-test format against a baseline established by official sources and experts in the fields of AI and strategic planning. The assessment includes metrics based on both self-reported learning and measurements of changes to participant responses to LO-related questions post-gameplay. The experiment found a strong correlation between the measured learning and participants' self-reported learning, and both metrics confirm that Obsolescence achieves its educational goals. This research includes the steps necessary to utilize the assessment methodology and presents recommendations both for Obsolescence and for future research in the field of educational game assessment.

15. SUBJECT TERMS
Artificial Intelligence, Serious Games, Game-Based Learning, Testing and Evaluation, Pedagogy

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Mark Reith, AFIT/ENG
U	**U**	**U**	**UU**	130	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-6565, ext 4603 (mark.reith@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18