Human Factors in Robots, Drones and Unmanned Systems, Vol. 93, 2023, 199–218
https://doi.org/10.54941/ahfe1003948

To Shoot or Not to Shoot? Human, Robot, & Automated Voice Directive Compliance With Target Acquisition & Engagement

Giovanna Camacho, Matthew Bolton, Joseph Loggi, Kallia Smith, Emmett Rice, and Tariq Iqbal

University of Virginia, Charlottesville, VA 22903, USA

ABSTRACT

The Army's Optionally Manned Fighter Vehicle (OMFV) program seeks to, "...operate with no more than two crewmen" (Congressional Research Service, 2021), but currently uses four individuals: driver, gunner, commander, and ammo handler. This study sought to investigate how automated teammates affects war fighters within the tank. To achieve our research objective, we performed a human subjects' study under IRB ID 5734, from the University of Virginia. This experiment was a mixed measures design as all participants were tasked to take directives from three entities, but half of the participants were given directives by a female voice while the other half were given a male voice from all entities. Participants were tasked to take commands from a human, NAO robot, and a computer automated voice while deciding on whether to fire upon armed robots, a swarm of drones, or a single drone. They engaged targets by use of a computer mouse. Participants were instructed that the commands given to them might not be correct and it was upon their judgment if the target was indeed a necessary target. The entire experiment took approximately 30 minutes in total as there were 54 iterations where participants were given 20 seconds to respond with a click totalling 18 minutes that left them with 12 minutes where they completed a demographic survey, NASA TLX, SART, gave subjective feedback, and were briefed and debriefed. Data was analyzed using mixed linear model ANOVAs. Overall, army participants preferred instruction from a human. Less experienced users completely ignored all directives given and proceeded to engage as they saw fit. Individuals given directives from the computer had lower accuracy and situational awareness (SA) scores. Individuals directed by the computer had lower workload scores than they did being directed from a human, but higher workload scores than when directed by the robot. Human directed participants had a higher workload and situational awareness scores. Higher accuracy scores were seen in target acquisition, but not in target engagement for individuals directed by the robot. Participants receiving directives from the robot had the lowest workload score on average and had a moderate SA score. Participants never looked at the robot during the experiment once it began, as they were task saturated with their vision fixated on their targets while listening for commands. Participants felt the least workload from the robot but moderate frustration with the robot and the highest frustration with the computer automated directives. There were significant differences found between the computer and robot directives when it came to SA ($F(2,26) = 3.48$, $p < .046$, $np2 = .211$). There were also significant differences between the accuracy target engagement scores of the beginner and experienced participants ($F(2,22) = 3.83$, $p < .037$, $np2 = .258$). There were no differences in how participants responded between male/female directives voices. Furthermore, the robot utilized did not show preference to male/female directives either to initiate mission directives. Ultimately, data produced in this study will help understand how to best facilitate operator performance with or without Human Automated Teammates.

Keywords: Human robot interaction (HRI), Human agent teaming (HAT) & situation awareness

INTRODUCTION

Contemporary initiatives from the United States Army and the Department of Defense (DOD) have sought to update and modernize vehicles such as the M-1 Abrams Tank, The M-2/M-3 Bradley Fighting Vehicle (BFV), and the M-1126 Stryker Combat Vehicle (Feickert, 2016). These infantry fighting vehicles (IFVs) serve as the strength of the Army's Armored Brigade Combat Teams (ABCTs) and Stryker Brigade Combat Teams (SBCTs) and have been in service since 1980. These older weapon systems require modernization to not only maintain operation, but also to maintain effectiveness with the kinetic advancing battlefield. Part of the DOD's modernization initiative is the Next Generation Combat Vehicle (NGCV) program. This program specifically focuses on requirements such as, being optionally manned, operated remotely, and utilizing artificial intelligence with certain tasks to assist crewmembers. Part of being successful in this initiative is recognizing both where the human operator and NGCV fall short and what actions can be taken to enhance their effectiveness. Currently the M-1A2 can "acquire targets 45% faster and hand off targets 50-75% faster, thus giving it a percent to hit on evasive targets that is 80% better than an M-1A1 '(Feickert, 2016). These improvements are critical to continued mission success, however, operators are faced with overwhelming workloads, task overload and situational awareness deficiencies. Along with the human operator's predisposed difficulty of maintaining vigilance, increased mental workloads have been shown to reduce the size of the operator's visual field due to it directly affecting the operator's situational awareness (Rantanen & Goldberg, 1999). During daily missions, fast flowing information relayed to a human operator can lead to skewed actions when a human finds one set of information more salient above the rest causing them to overlook alternative data. This is referred to as anchoring heuristic bias (Wickens, 2005). Task completions may be affected by communication constraints through latency between machine control inputs and observable changes in sensor feeds (Rastogi, 1996). Human operators also have genetically predisposed characteristics of spatial ability which could stand to impact performance if their spatial ability is not high (Chen et al., 2008).

Outside of increased training to navigate through errors of the human operator, autonomy in the NGCV offers key solutions to enhance mission success. Autonomy of the NGCV allows for less war fighters to be in harm's way. High amounts of mission data can be analyzed in large volumes allowing for a full assessment of courses of actions, which will increase the quality and speed of decisions in time-critical operations. Complex missions can be completed utilizing multimodal assets as the autonomous systems can analyze all data assets simultaneously. The autonomous system can still operate when communication is intermittent, thereby allowing the possibility for continued missions for longer durations. More dangerous missions can occur without the loss of human lives (Neilsen & Ruth, 2016).

Automation must allow for the flexibility of adjustments as new intel is relayed to the leadership operating the system. The key instruction to facilitating this transition is defining the reality that, "…autonomy results from

delegation of a decision to an authorized entity to take action within specific boundaries…systems governed by prescriptive rules that permit no deviations are automated, but they are not autonomous" (Neilsen & Ruth, 2016). It is at this juncture that the question arises of how much automation should be considered when integrating humans and automated systems? There lies an understanding that, "automation does not exist in an all or none fashion but can be implemented at various levels" (Endsley, 1997).

The Army's OMFV program currently states that the tank, "...should eventually operate with no more than two crewmen"(Congressional Research Service, 2021). Currently the crew has four individuals: driver, gunner, commander, and ammo handler. The position to be considered for removal and automated is the ammo handler. This study sought to compare receiving target acquisition and confirmation between: a human teammate, computer automated voice, and a NAO robot. The objectives of this study would facilitate the comparison of participants receiving guidance from these entities and allow results that could define better practices of automation that would enhance the war fighter's capabilities within the tank.

The objectives of this study were as follows:

- Facilitate a realistic simulated environment where a beginner, intermediate, and experienced war fighter are assessed on target acquisition and engagement with the help of a teammate (human, computer automated, and robot).
- Gain an understanding of how the change in teammate affected the time, accuracy, mental demand, physical demand, temporal demand, performance, effort, frustration, and situational awareness of the end user in their target acquisition and engagement.
- Analyze data to consider what future studies should be considered in the human agent teaming interactions that will facilitate the movement toward automated systems.

The hypotheses that are being tested are as follows:

Hypothesis 1: If the user is more experienced, then they will favor the instruction of the human agent. If the user is less experienced, then they will favor the instruction of the robot agent (Sung-en Chien et al., 2019).

Hypothesis 2: If the participants are given directives from the computer automated system, then the target acquisition and engagement would be faster with a higher accuracy. However, these participants would have a higher workload score for the NASA Task Load Index (TLX) and lower situational awareness SART score.

Hypothesis 3: If the participants are given directives from a human the target acquisition and engagement would be slower with lower accuracy. However, these participants would have a lower workload score for the NASA TLX and lower situational awareness Situational Awareness Rating Technique (SART) score.

Hypothesis 4: If the participants are given directives from a robot the target acquisition and engagement would be faster with higher accuracy. However, these participants would have a moderate workload score for the NASA TLX and a moderate situational awareness SART score.

The sub-questions of interest are as follows:

Does replacing the human with a robot that still shows nonverbal communication such as body language, gaze cues, etc. more influential for making decisions for soldiers than computer automated voice?

Can a robot accomplishing the same task of identifying a target as a human be confidently received at the same level as the human teammate?

This study overall found unexpected results and supported preconceived notions in age disparities of users (Chien et al., 2019). Army participants preferred instruction from a human. Less experienced users completely ignored all directives given and proceeded to engage as they saw fit.

Individuals given directives from the computer had lower accuracy scores and lower situational awareness (SA) scores. Individuals directed by the computer had lower workload scores than they did being directed from a human, but higher workload scores than when directed by the robot.

Human directed participants overall had a higher workload score and higher situational awareness. Higher accuracy scores were seen in target acquisition, but not in target engagement for individuals directed by the robot.

Participants receiving directives from the robot had the lowest workload score on average and had a moderate SA score. Participants never looked at the robot during the experiment once it began as they were task saturated with their vision fixated on their targets while listening for commands.

RELATED WORK

The U.S. Army Combat Capabilities Development Command (DEVCOM) first began working towards the Modernization Priority NGCV 2019 standing up their Human Autonomy Teaming (HAT) research program (DEVCOM, 2023). The introduction of the 2018 U.S. Army Modernization Strategy Report initiated by congress in 2018, dictated the modernization priorities of a, "multi-domain force by 2035 " (U.S. Army, 2019). In aligning with this goal, DEVCOM has moved forward in research to analyze Human Agent Teaming (HAT) methods from various perspectives to understand the capabilities, performance, risk, and decisions made during missions sets from Soldiers working in conjunction with autonomous systems. The research organization has found that the call for a, "reduction in crew size, will require effective teaming with emerging technologies, such as AI, autonomy, and robotics, in order to succeed" (DEVCOM, 2023).

This project was inspired through the work of DEVCOM and companies such as Pratt Miller looking to understand the complexities of the HAT as they move forward in their development methods of creating the updated NGCV. Various journals presented by DEVCOM have outlined SA with HAT effectiveness in multitasked scenarios, as well as the range of transparency of information between the HAT being guided by trust (Barnes et al., 2019; Chen et al., 2018; Chen et al., 2013). Within these papers the interaction between end user and the autonomous agents are tested, but testing the differences in compliance was not expanded upon such as suggested by Haring et al., 2021, in a less kinetic trivial visual search task where they compare a

Nao robot, Roomba robot, Baxter, and a real human robot. Both DEVCOM and Haring et al., also considered trust, anthropometric ratings, SA, workload, accuracy, time, and personality traits. However, moving forward this pilot study focused on accuracy, time, workload, and situational awareness to focus on the performance of the end user. The goal of this experiment was to not fixate on analyzing the potential causation that predisposed attributes of a participant may have on their performance. Training can effectively level out these predispositions, it is the performance that needs to be evaluated for the Soldiers affected by these changes.

METHODS/EXPERIMENTAL DESIGN

Participants were gathered from three groups: 5 random students at the University of Virginia (UVA), 5 ROTC students from UVA, and 4 active-duty Army Soldiers affiliated with UVA. The participant break down was as follows:

Gender:	4 females and 10 males
Age:	19–43
Handedness:	2 left handers
Use of Glasses:	5
Hours of Sleep:	4–10 hours
Years of Military Experience:	0–20
Years of ROTC Experience:	0–4
Education:	All had some college
Majors:	Economy, Foreign Affairs, Commerce, Cognitive Science, Clinical Pastoral Education Program, Commerce Management, Mechanical Engineering, Civil & Environmental Engineering, Public Policy, Human Biology and History

All had gaming experience with two participants listing Call of Duty in common

Only three individuals had any remote-control joy-stick type vehicle usage

All participants were asked for their voluntary participation and given consent forms where they were told they could leave the experiment at any time should they not wish to continue. This study took place in the Human Factors Engineering Lab in the basement of Olsson Hall at the University of Virginia. Participants utilized a Microsoft Surface Pro, Wom Mouse, and utilized a PowerPoint as a simulation. The set-up can be seen in Figure 1.

This experiment was a mixed measures design as all participants experienced directives from three entities, but half of the participants were given directives by a female voice while the other half were given a male voice from all entities. Participants were tasked to take commands from a human, NAO robot (see APPENDIX A for algorithm), and a computer automated

Figure 1: UVA student, ROTC and military personnel undergoing the simulation.

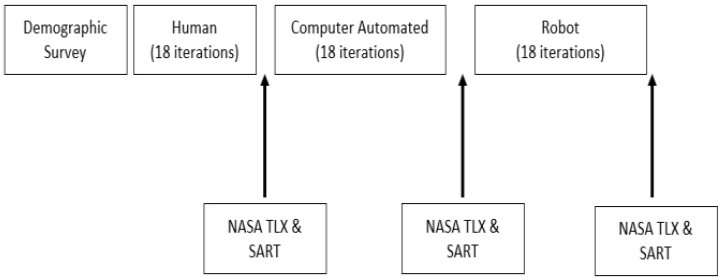


Figure 2: Schematic of experimental flow.

voice while deciding on whether to fire upon armed robots, swarm of drones, or a single drone (all enemy entities were referred to as Technovians). They engaged targets by use of a mouse. Participants were instructed that the commands given to them might not be correct and it was upon their judgment if the target was indeed a necessary target. The entire experiment took approximately 30 minutes in total as there were 54 iterations where participants were given 20 seconds to respond with a click totalling 18 minutes that left them with 12 minutes where they completed the demographic survey, NASA TLX, SART, give subjective feedback, and were briefed and debriefed.

The overall flow of the experiment looked like this with each block (Human, Computer Automated & Robot) of directives presented to each participant in a different order to avoid and potential for carryover effects in the data.

Data Capture & Analysis Method

Data was collected both by a computer program and from the participant. The computer program was a coded PowerPoint (see APPENDIX B) that calculated the time an individual took to shoot along with the accuracy of their shot. This accuracy was also verified by an individual monitoring the individual during the scenario along with another individual recording subjective feedback from the participant. The data collected from the participant was general demographic data (see above). This data was stored within the lab and collected by pen and paper to avoid any potential for loss of data electronically. The NASA TLX and SART information was collected three times throughout the experiment after each block of directives (human, robot, and computer automation) as well as the subjective reasoning for engaging or not

engaging each target throughout the simulation. The NASA-TLX is a six-factor index (Hart & Staveland, 1988) which divides workload into: mental demand, physical demand, temporal demand, performance, effort, and frustration. The test-retest rating was good at .83 (Hart & Staveland, 1988). The SART is a ten-factor index which considers the attentional domains of demand, supply and understanding. This test has a medium rating of effectiveness (Selcon et al., 1989) and we acknowledge that this test has validity issues (M. L. Bolton, 2022). The SART was a better option than the Situation Awareness Global Assessment Technique, as the goal was to maintain a higher level of stress amongst our participants. Keeping the missions as realistic as possible would not have occurred if there were freezes in scenarios to acquire data through probing questions.

All data was analyzed by utilizing Statistical Package for the Social Sciences (SPSS) with their mixed linear model capabilities. Statistical analyses of the dependent variable directive with three levels (computer, human and robot) along with variances in utilizing a male or female voice was evaluated. The independent variables (NASA TLX, SART, accuracy and time) were evaluated with three one-way repeated measure ANOVAs followed up with post hoc t-tests. This data will be stored within the lab for five years as dictated by IRB protocol.

RESULTS

The metrics of performance for this study were time and accuracy. All participants completed the mission sets in less than the 20 second threshold, but the computer system mixed up which times corresponded to mission sets so analyzing by mission set was not possible. Accuracy was broken down into two separate metrics. The first metric was accuracy with target acquisition in correctly identifying the enemy as seen in Figure 3.

Target Acquisition was 98% up to 100% for all participants regardless of directives given. Robot directives were compatible in accuracy with human

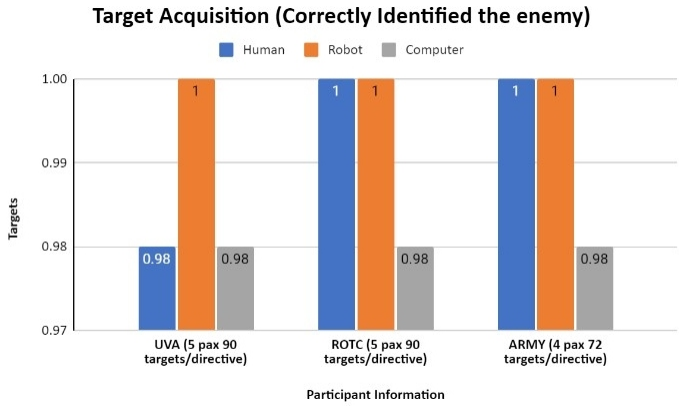


Figure 3: Target acquisition accuracy of correctly identifying the enemy across all participants by directive.

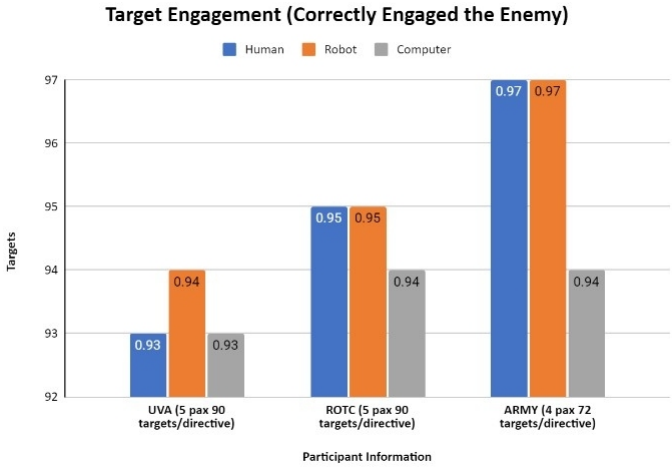


Figure 4: Target engagement accuracy of correctly engaging the enemy across all participants by directive.

directives for both ROTC and Military participants, although the robot directives surpassed the human directive accuracy for UVA students as one UVA participant misjudged a human directive given. The computer directives for each participant caused a miss of at least one target regardless of the type of participant.

The second metric was accuracy with target engagement in correctly identifying the enemy as seen in Figure 4.

Target Engagement was 93% up to 97% for all participants regardless of directives given. This was a metric that showed the ability of an individual to appropriately engage the enemy utilizing their judgment regardless of the guidance of the different mediums utilized to give them directives. Out of all participants, a 3rd year ROTC student and 13-year Army Soldier scored perfect for target acquisition and appropriate engagement of the enemy. Army and ROTC participants waited for full commands prior to firing. UVA students ignored directives with only one student listening to a full command when uncertain in engaging targets. All Army participants verbally relayed that they recognized not to shoot, yet some engaged anyway causing them to incorrectly engage the enemy. One Army participant correctly identified the enemy as a Technovian Soldier, however improperly engaged the enemy as this Technovian Soldier was unarmed.

There were no differences in target accuracy with female or male directives given as the mistakes made were made equally between male directive voice and female directive voice.

The overall averages of NASA-TLX scores among directive medium ranged from 20–27 which is a medium workload score that can be seen in Figure 5.

These scores show that the overall workload scores for the robot are at the highest while the human score was the lowest with computer directives falling slightly below the robot workload score.

Figure 6 shows the breakdown of the NASA TLX Workload factors by Directive Medium.

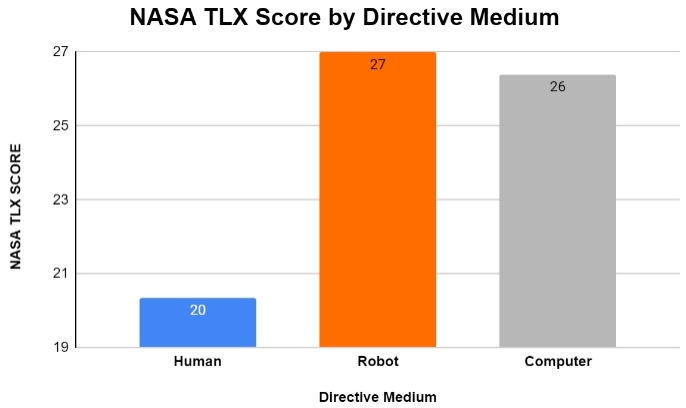


Figure 5: Overall NASA TLX score in response to directive medium given to participants.

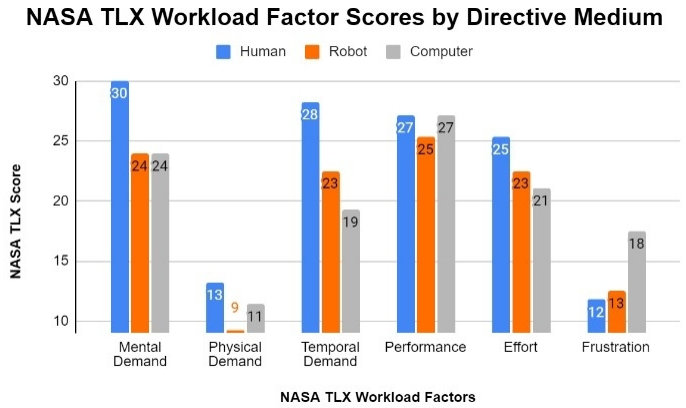


Figure 6: Breakdown of average NASA TLX workload factor scores by directive medium.

When breaking down the NASA TLX scores into components the mental, physical, temporal, performance, and effort component, scores were higher for human directives given. The only component score that was lower for the human workload scores was frustration which was lowest for the participants at a score of 12 who received human directives but highest for participants at a score of 18 who received computer directives. The frustration score for the robot was 13.

The overall averages of NASA -TLX scores by participant group ranged from 13–33 which is a medium to somewhat high workload score that can be seen in Figure 7.

The UVA participants had the highest workload scores with the ROTC participants falling just below that and then the Army participants with the lowest workload score. The computer workload was the lowest overall with the robot score being the next highest and the human workload score being the highest workload score amongst the participant groups.

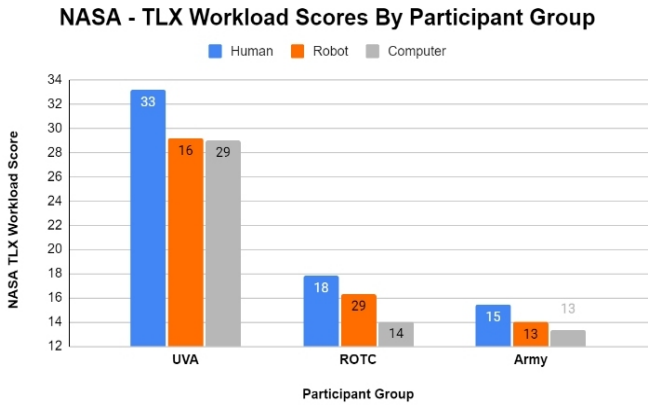


Figure 7: NASA TLX workload score by participant group.

The NASA TLX Average Workload Scores did not vary much between male and female voice directives given by various mediums although across the board female directives seemed to result in a slightly higher workload as seen in Appendix C Figure 8. The major difference shown for the Army participant in voice directives given was because one participant receiving male directives rated workload as zero across all factors. Overall, no significant differences were found relating to male or female voice directives.

The overall average of SA scores was on the low end as a score of 63 is the highest attainable score. The SA scores are shown in Figure 9.

The SA for participants receiving the human directives was highest followed by the robot and then the computer.

Breaking down the SA scores into their subcomponents of Attentional Demand, Supply, and Understanding is seen in Figure 10.

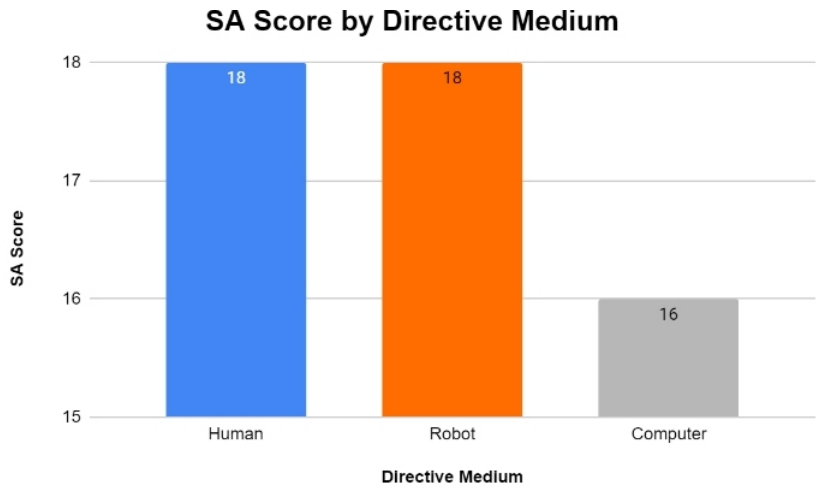


Figure 9: Overall SA scores in response to directive medium given to participant.

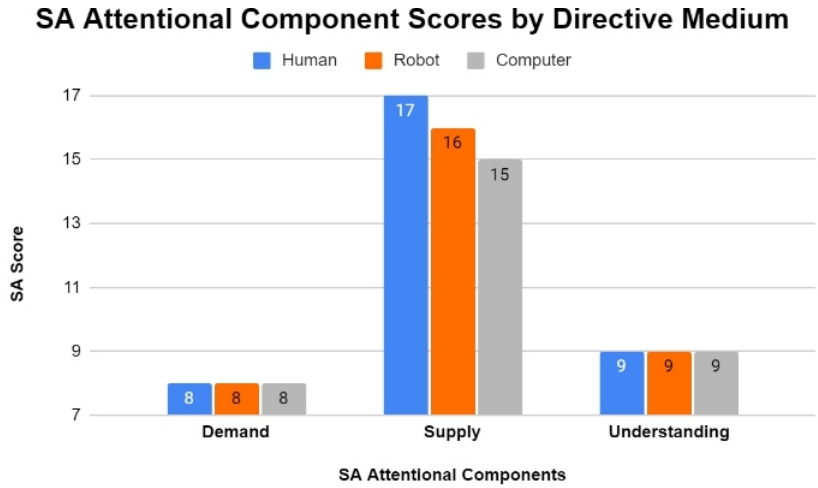


Figure 10: Breakdown of average SA attentional component scores by directive medium.

The highest possible SA score for demand when assessing scoring is 21 (participants scored rather low), for supply is 28 (participant scores mid-range), and for understanding is 14 (participants score on the higher end). The major difference in SA scoring occurred at the supply of attention with the human directive requiring the highest supply followed by the robot and then the computer directives.

Figure 11 shows the differences in SA between the participant groups.

Here you can see that the ROTC participants had the highest level of SA with human and robot directives being equally effective in attaining the participants SA. The computer had a lower SA score for military participants.

The female and male directives amongst the various mediums are shown in Appendix D Figure 12. Looking over this data the SA scores between female

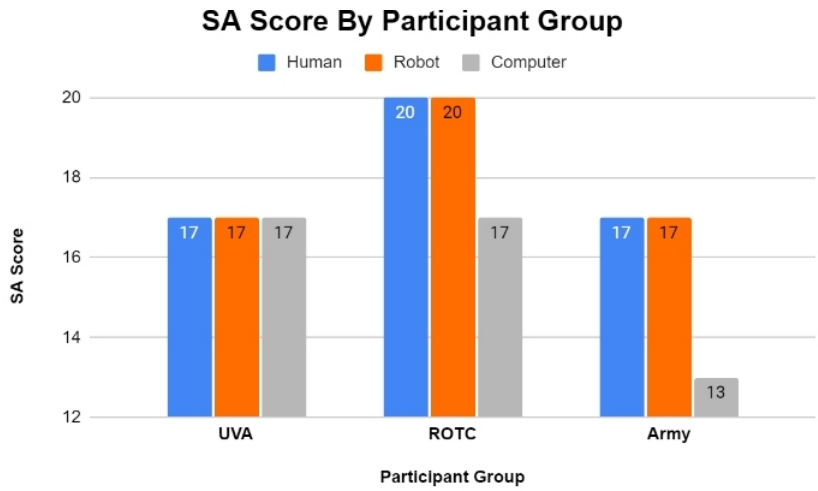


Figure 11: SA score by participant group.

and male did vary between directive medium and participant group but did not prove to be significant variations in the data set.

Three one-way repeated measure ANOVA tests were completed to assess if there was a change in score with the NASA TLX/SA accuracy – target engagement over the three directive conditions (computer, robot, or human). The significance for change in scores in SA was found between computer and robot directives. Scores of SA were compared for the various directives given. There was a main significant effect on the mean SA scores between computer and robot directives $F(2,26) = 3.48$, $p < .046$, $np2 = .211$. There was a large effect on the SA score from the computer and robot directives given. A paired samples t test was conducted to evaluate the impact of the computer and robot directives given on the SA scores and there was a statistically significant decrease in SA scores between the robot directives ($M = 17.86$, $SD = 4.55$) to computer directives ($M = 15.64$, $SD = 4.43$), $t(13) = 13.21,14.67$, $p < .001$ (two-tailed). The mean decrease in SA scores was 2.22 with a 95% confidence interval ranging from 13.08 to 18.21 for the computer and 15,22 to 20.49 for the robot. The partial eta squared statistic was .93 for the computer and .94 for the robot indicating a large effect size. To evaluate the SA scores further to sort out the underlying component of the SA score causing the significant difference, a paired sample t test of SA scores was completed. The correlation of the subcomponent scores of SA was too high to sort out any further differences that could explain the differences in mean scores.

Another question assessed was if there was a change in score accuracy-target engagement over three different groups utilized (UVA, ROTC, and military). There was a main significant effect on the mean accuracy target engagement scores between UVA students and military scores $F(2,22) = 3.83$, $p < .037$, $np2 = .258$. The pairwise comparison further evaluated the UVA student ($M = 16.83$, $SD = .39$) and Military ($M = 17.33$, $SD = .65$) engagement scores to be $p = .026$.

Nine t -tests were conducted to evaluate the difference in male/female directives affecting performance in scores for the NASA TLX, SA, and accuracy - engagement scores. No significant differences were observed.

Additional findings with regards to the NAO robot utilized, assessed the effectiveness of the robot taking directives from a female and male end user to engage in the appropriate mission set. Out of 252 missions NAO faulted 96 of them making it 38% ineffective. The NAO robot was faulty to both male and female voice recognition 48 times evenly. Thereby making the voice recognition equally ineffective regardless of female or male voice.

DISCUSSION

The results of this pilot study proved to be extremely useful in helping to analyze metrics that mattered most during mission sets. The first metric that was realized to be ineffective within this study was time. Each participant chose their method in handling the stressful situation presented to them and was able to meet the twenty second threshold. The ability to recognize the target was fast but the individual takes time to engage to reach the target should not be an issue.

The importance of training was highlighted in this experiment through the various participant groups. One scenario utilized a M2 .50 caliber machine gun to engage enemy within 1,000 meters of friendlies. Trained individuals recognized this was a moment when they should not engage when technology told them to. In this instance, the technology would be programmed with maximum range, however an individual would be able to distinguish the maximum effective range and understand fratricide to be a possibility in this complex scenario. Training helps as individuals move from their forebrain in thinking to their midbrain when they are angry or frightened. The midbrain is similar to animalistic thinking, so overcoming this mindset requires operant conditioning when individuals are trained on how to respond to a stimulus repetitively. With enough training, individuals can respond appropriately overcoming the innate functionalities of their pre-programmed midbrains. Should variances in directives occur, training should be moved to the forefront to overcome any differences within the directives.

Overall, the participants felt the least workload from the robot but moderate frustration with the robot and the highest frustration with the computer automated directives.

There were significant differences found between the computer and robot directives when it came to SA, but no definitive understanding of what this meant since the components of the SA scores were highly correlated. There was also significant differences between the accuracy target engagement scores of the UVA students and military participants. There were no differences in how participants responded between male/female directives voices. Furthermore, the robot utilized did not show preference to male/female directives either to initiate mission directives.

The workload scores were lowest for the human directive overall, yet across components the human directive had the highest workload by factor. The curious finding within the workload data was that the only factor that the human directive had the lowest component in across factors of workload was frustration. This could imply that individuals find technology to be more frustrating than working with a human counterpart. The robot was only rated slightly higher in frustration than the human directive with the computer being rated the highest. The workload scores seemed to further support the need for training, as the military participants had lower workload scores overall with the ROTC participants following closely behind. UVA students found the mission set to have the highest workload since they have not been exposed to similar scenarios.

Situational awareness scores were ranked the same between the robot and the human directives given. The attentional supply component of the SA score seemed to have variation amongst directive mediums. The significant differences found between the computer and robot directives could be attributed to this factor, although the strong correlation within the SA components did not have supportive evidence to make this finding conclusive.

This pilot study was useful in analyzing metrics that directly affect individual performance. The SA metric is an area that needs to be further understood and evaluated to truly understand what is occurring amongst the components of the score evaluation. Workload and accuracy engagement

need to be further evaluated as well to help further manage performance expectations. Overall, training of the various participant groups explicitly showed prominence in overcoming any directive differences.

FUTURE WORK

Future work should be completed in this study with more participants to truly analyze the differences in performance. Current work in this field explicitly focuses on innate attributes such as "Trust," of individuals not realizing that these values take a backseat when placed under pressures in the life-or-death situations Soldiers face. Furthermore, Trust is not a subjective measure that can adequately be tied back to how it directly affects performance under duress. Soldiers will utilize equipment that effectively helps them complete their missions so they can come home.

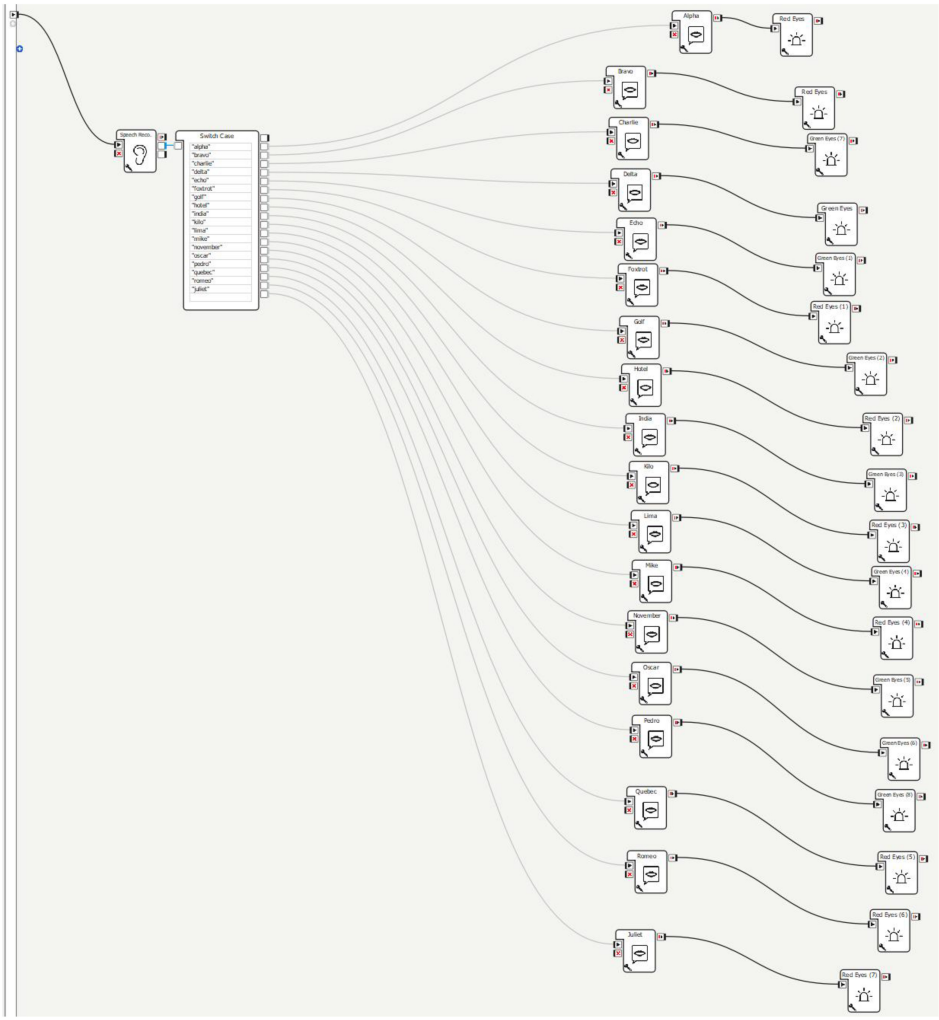
Understanding the differences in performance can help to pinpoint areas that could be evaluated for improvements in creating a better HAT. Analyzing this pilot experiment with real human targets in more realistic simulations such as unreal game engine, virtual reality modalities, and in person scenarios at gunnery ranges would also be opportunities to consider for, "improved data collection and analysis" that DEVCOM is in search of (DEVCOM, 2023). Completing more in- depth training on weapon systems and laws of engagement with future participants would also allow for better understanding of accuracy in selection of target engagements.

Furthermore, research in considering a flashing border on a HUD to help individuals know when to engage targets rather than just listening to auditory commands could help in scenarios where the environment is either too loud for oral communication or needs more stealth in communications.

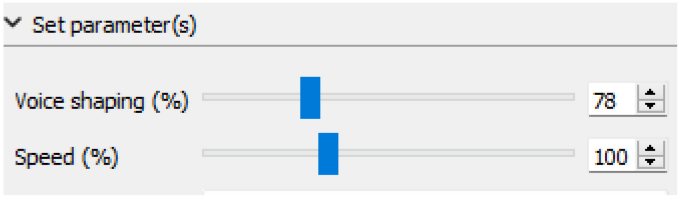
Overall, the changing of human teammates to robots and effectively phasing out all humans on engagement of the enemy does beg for both psychological and ethical questions that should be further researched. The act of killing as researched by Grossman (2009), "On Killing" discusses the "intimacy and psychological impact," of taking another human life. By replacing a human teammate who will share the burden of engaging the target? By replacing human beings with all autonomous entities where does the killing end? Glenn Gray, a philosopher from WWII described the desensitization at maximum range (range at which an individual is unable to perceive individual victims without technological assistance) phenomenon stating, "Many a pilot or artilleryman who has destroyed untold numbers of terrified noncombatants has never felt any need for repentance or regret" (Grossman, 2009).

APPENDIX A ROBOT CODED

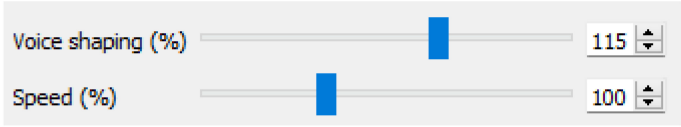
The robot will listen for the military alphabetic letter that corresponds to the mission. Upon hearing the specific alphabet letter stated the robot states the mission of engagement and accompanies the information with either red or green eyes helping to relay to the participant whether or not they should engage.



Male Robot Voice



Female Robot Voice



APPENDIX B MASTER GLOBAL DIRECTIONS FOR CODED POWERPOINT

```
'Option Explicit

   Type POINTAPI
   Xcoord As Long
   Ycoord As Long

End Type

Public Declare PtrSafe Function GetCursorPos Lib "User32" (lpPoint As
POINTAPI) As Long

Public Declare PtrSafe Function GetSystemMetrics32 Lib "User32" Alias
"GetSystemMetrics" (ByVal nIndex As Long) As Long

Public Const OutFile As String = "datalog.txt"

Function GetScreenWidth()
      GetScreenWidth = GetSystemMetrics32(0)
End Function

Function GetScreenWidth()
      GetScreenWidth = GetSystemMetrics32(1)
End Function

Function GetSlideTime()
      GetSlideTime = SlideShowWindows(1).View.SlideElapsedTime
End Function

Function MouseInShape(Slide As Object) As Boolean
      Dim Shape As Object
      Dim MouseCoord As POINTAPI
      Set Shape = FindRectangleShape(Slide)

      GetCursorPos MouseCoord
      newmousex = MouseCoord.Xcoord * ActivePresentation.PageSetup.
SlideWidth / GetScreenWidth()
      newmousey = MouseCoord.Ycoord * ActivePresentation.PageSetup.
SlideHeight / GetScreenHeight()

      If (newmousex >= Shape.Left And newmousex <= Shape.Left +
Shape.Width) And (newmousey >= Shape.Top And newmousey <= Shape.Top
+ Shape.Height) Then
      MouseInShape = True
   Else
      MouseInShape = False
```

```
    End If
End Function

    Function FindRectangleShape(Slide As Object) As Object
        For Each Shp In Slide.Shapes
            If Shp.AutoShapeType = msoShapeRectangle Then
                Set FindRectangleShape = Shp
            End If
        Next
End Function

    Sub RecordData(Slide As Object, TargetClicked As Boolean, TimeTaken
As Double)
        Dim FileNum
        FileNum = FreeFile
        Open Application.ActivePresentation.Path & "\" & OutFile For
Append As FileNum
        Print # FileNum, "TimeStamp: " & Now() & ", Slide: " &
Slide.Name & ", TargetClicked: " & CStr(TargetClicked) & ", Time(s): "
& CStr(TimeTaken)
    Close FileNum
End Sub

Sub CollectInput(Slide As Object)
    RecordData Slide, MouseInShape(Slide), GetSlideTime()
    ActivePresentation.SlideShowWindow.View.Next
End Sub
```

APPENDIX C NASA TLX AVERAGE WORKLOAD SCORES BY PARTICIPANT GROUP IN RESPONSE TO MALE/FEMALE VOICE DIRECTIVES

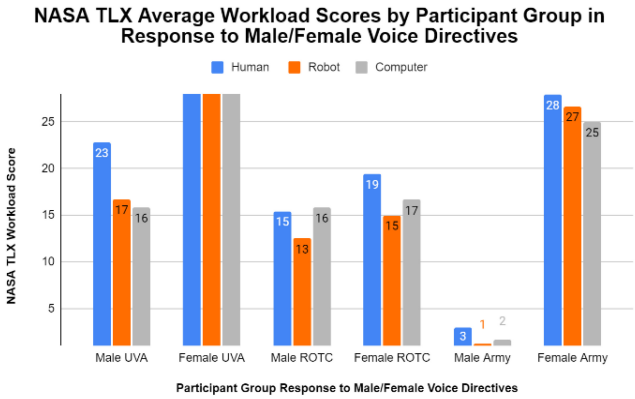


Figure 8: NASA–TLX workload score by participant group in response to male/female voice directives.

APPENDIX D SA AVERAGE SCORES BY PARTICIPANT GROUP IN RESPONSE TO MALE/FEMALE VOICE DIRECTIVES

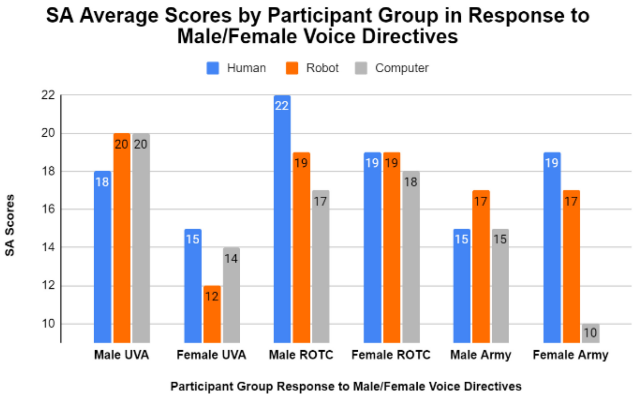


Figure 12: SA score by participant group in response to male/female voice directives.

ACKNOWLEDGMENT

This paper was created for a Human Robot Interaction class project at the University of Virginia.

REFERENCES

Army Publishing Directorate – Details Page, "TC 3–20. 31–4." https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB_ID=105365.

Ashish, Gawali. Answered questions on normalization of accuracy graphs. 29APR23.

Barnes, Michael & Elliott, Linda & Wright, Julia & Scharine, Angelique & Chen, Jessie. (2019). Human-Robot Interaction Design Research: From Teleoperations to Human- Agent Teaming Human-Robot Interaction Design Research: From Teleoperations to Human-Agent Teaming.

Bolton, Matthew. Interviewed multiple times to double check experiment design, ask questions regarding details of experiment, and receive help on building the powerpoint code for the simulation. 23MAR23.

Chen, C., J. Y. and Barnes, M. J., Qu, Z, & Snyder, M. G. (2010). Roboleader. An Intelligent Agent for Enhancing Supervisory Control of Multiple Robots.

Chen, Jessie & Lakhmani, Shan & Stowers, Kimberly & Selkowitz, Anthony & Wright, Julia & Barnes, Michael. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. Theoretical Issues in Ergonomics Science. 19. 259–282. 10.1080/1463922X.2017.1315750.

Chien et al., "Age Difference in Perceived Ease of Use, Curiosity, and Implicit Negative Attitude toward Robots Age Differences in Attitudes toward Robots." Age Difference in Perceived Ease of Use, Curiosity, and Implicit Negative Attitude toward Robots, https://dl.acm.org/doi/fullHtml/10.1145/3311788. 2019.

DEVCOM. "Essential Research Programs." DEVCOM Army Research Laboratory, https://www.arl.army.mil/what-we-do/hat/. 2023.

Du Hemin, Wang Ting. Study on Man-machine Design of Armoured Vehicles Operation Space. MATEC Web Conference (2018). 10.1051/matecconf/201817604004.

Endsley, Mica R. "Level of Automation: Integrating Humans and Automated Systems." Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 41, no. 1, 1997, pp. 200–204., doi: 10.1177/107118139704100146.

Feickert, Andrew. "The Army's M-1 Abrams, M-2/M-3 Bradley, and M-1126 Stryker: Background and Issues for Congress." Congressional Research Service, Https://Fas.org/Sgp/Crs/Weapons/R44229. Pdf, 5 Apr. 2016, Congressional Research Service.

Grossman, Dave. On Killing: The Psychological Cost of Learning to Kill in War and Society. Little, Brown and Co., 2009.

Haring Kerstin S., Satterfield Kelly M., Tossell Chad C., et al. Robot Authority in Human - Robot Teaming: Effects of Human LIkeness and PHysical Embodiment on Compliance. Frontiers of Psychology. VOL. 12, 2021. 10.3389/fpsyg.2021.625713.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load INDEX): Results of empirical and theoretical research. Advances in Psychology, 139–183. doi: 10.1016/s0166-4115(08)62386-9

Iqbal, Tariq. Received guidance on what our study needed to include such as additional participants and adding the experimental directive of male and female voices.23MAR23

J. Khurshid and Hong Bing-rong, "Military robots - a glimpse from today and tomorrow," ICARCV 2004 8th Control, Automation, Robotics and Vision Conference, 2004., Kunming, China, 2004, pp. 771–777 Vol. 1, doi: 10.1109/ICARCV.2004.1468925.

J. Y. C. Chen, S. Quinn, J. Wright, M. Barnes, D. Barber and D. Adams, "Human-agent teaming for robot management in multitasking environments," 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 2013, pp. 103–104, doi: 10.1109/HRI.2013.6483522.

Kempinski, Bernard, and Christopher Murphy. Technical Challenges of the U. S. Army's Ground Combat Vehicle Program. Washington, D. C: Congressional Budget Office, 2012. Print.

Michael Watson, Christina Rusnock, Michael Miller, and John Colombi. 2017. Informing System Design Using Human Performance Modeling. Syst. Eng. 20, 2 (March 2017), 173–187. https://doi.org/10.1002/sys.21388

M. L. Bolton, E. Biltekoff and L. Humphrey, "The Level of Measurement of Subjective Situation Awareness and Its Dimensions in the Situation Awareness Rating Technique (SART)," in IEEE Transactions on Human-Machine Systems, vol. 52, no. 6, pp. 1147- 1154, Dec. 2022, doi: 10.1109/THMS.2021.3121960.

"NASA Task Load Index (NASA TLX)." NASA Task Load Index (NASA TLX) | HP Repository, https://ext.eurocontrol.int/ehp/?q=node%2F1583.

Nielsen, Paul and Ruth A. David. "Defense Science Board Summer Study on Autonomy." 2016, doi:10.21236/ad1017790.

Rantanen, E., & Goldberg, J. (1999). The effect of mental workload on the visual field size and shape. Ergonomics, 42, 816–834.

Rastogi, A. (1996). Design of an interface for teleoperation in unstructured environments using augmented reality displays. Unpublished master's thesis. University of Toronto, Canada.

Selcon, S. J. & Taylor, R. M. (1989). Evaluation of the Situational Awareness Rating Technique (SART) as a tool for aircrew systems design. Proceedings of the AGARD

AMP Symposium on Situational Awareness in Aerospace Operations, CP478. Seuilly-sur Seine, France: NATO AGARD.
"Situation Awareness Rating Technique (SART)." Situation Awareness Rating Technique (SART) | HP Repository, https://ext.eurocontrol.int/ehp/?q=node%2F1608#: ~:text=SART%20is%20a%20post%2Dtrial,a%20seven%20point%20rating%20scale.
Sung-En Chien, Li Chu, Hsing-Hao Lee, Chien-Chun Yang, Fo-Hui Lin, Pei-Ling Yang, Te-Mei Wang, and Su-Ling Yeh. 2019. Age Difference in Perceived Ease of Use, Curiosity, and Implicit Negative Attitude toward Robots. J. Hum.-Robot Interact. 8, 2, Article 9 (June 2019), 19 pages. https://doi.org/10.1145/3311788
"The Army's Ground Combat Vehicle Program and Alternatives." Congressional Budget Office, 2 Apr. 2013, https://www.cbo.gov/publication/44044.
The Army in Military Competition - Swagger Lume API. https://api.army.mil/e2/c/downloads/2021/03/29/bf6c30e6/csa-paper-2-the-army-in-military-competition.pdf.
The Army's Optionally Manned Fighting Vehicle (OMFV) - Congress. https://crsreports.congress.gov/product/pdf/IF/IF12094.
United States Army. https://api.army.mil/e2/c/downloads/2021/03/23/eeac3d01/20210319-csa-paper-1-signed- print-version.pdf.
U. S. Army Modernization Strategy - United States Army. https://www.army.mil/e2/downloads/rv7/2019_army_modernization_strategy_final.pdf. 2019.
Wickens, C. D. (2005, April). Attentional tunneling and task management. Paper presented at the 13th. International Symposium on Aviation Psychology, Dayton, OH.
Workload and Stress of Crews Operating Future Manned Vehicles. https://apps.dtic.mil/sti/pdfs/ADA463512.pdf.