THE DESIGN OF A MULTIMODAL DATA CAPTURE AND GESTURE RECOGNITION FRAMEWORK TO ENABLE FLUENT HUMAN-ROBOT COLLABORATION

(Technical Paper)

UNDERSTANDING CHOICE ARCHITECTURE IN THE AMERICAN HEALTHCARE SYSTEM

(STS Paper)

A Thesis Prospectus submitted to the
Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In partial fulfillment of the requirements of the degree
Bachelor of Science, School of Engineering

Evan Smith

Fall 2024

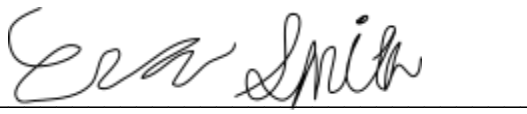
Technical Project Team Members

Camp Hagood

Sarah Naidu

Aramis Rolly

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signature  Date 11/8/24
Evan Smith

Technical Advisor: Tariq Iqbal, Department of Computer Science, Department of Systems Engineering

STS Advisor: Richard Jacques, Department of Engineering and Society

Introduction

From massive distribution centers to manufacturing facilities to our own homes, autonomous robots are becoming more prevalent and well-established in our working lives (Fairchild, 2021). However, autonomous robots aren't fully utilized by the United States Army in a warfare setting. This has the potential to unnecessarily put soldiers' lives at risk, especially during reconnaissance-based tasks. For autonomous robots to work alongside humans on the battlefield, humans must be able to communicate nonverbally with them due to the danger associated with verbal communication. One facet for enabling this nonverbal communication is gesture recognition. Gestures are widely used by all U.S. military branches, which would minimize the time and effort required to train soldiers to communicate with autonomous robots. With the following in mind, the focus of my capstone team's technical work is to enable fluent human-robot collaboration through gesture-based recognition.

It does not take a seasoned policy analyst to tell you that the American healthcare system is fundamentally broken. Only a quarter of Americans believe the current system works well, while the rest feel it requires fundamental changes or must be completely rebuilt (Schoen et al., 2013). Problems within the healthcare system include unaffordable care, limited access to insurance, difficulty understanding what insurance covers, difficulty navigating the system, low understanding of healthcare providers' recommendations, and a perceived inferior quality of care (Ducharme, 2023). While there is not a simple solution to improve American healthcare for everyone, it is important to understand how patients are currently able to interact with the healthcare system and how providers work with patients. My research will focus on understanding the implications and ethical considerations of choice architecture in American healthcare from the perspectives of both patients and providers.

Technical Discussion

The goal of our project is to develop a proof-of-concept gesture recognition framework that will be used in real-time on robots in a dynamic environment. The gestures we intend to recognize are some of the hand and arm signals used by the United States Marine Corps. After a comprehensive literature review of different state-of-the-art gesture recognition frameworks, we believe using a zero-shot visual language model (VLM) to large language model (LLM) gesture classification pipeline is the most promising approach. In real-time, the robot will take in a gesture and a predetermined prompt as an input for the VLM, which will generate a description of the gesture. Then, the generated description will be added to a predetermined prompt and sent to the LLM, which will provide the most likely gesture based on the prompt. The general approach for using these language models has been demonstrated to command robots in real time, although it does not currently utilize gesture recognition (Benjdira et al., 2023). In future iterations of our work, this approach may also allow for the robots to form a reasonable hypothesis about the person's intentions with or without explicit gestures and respond accordingly (Shi et al., 2024).

Our main obstacle to developing an effective VLM to LLM pipeline is developing effective prompts for both language models such that the pipeline will often produce correct classifications. An incorrect classification on the battlefield could lead to the robot conducting the wrong order, which can be disastrous on the battlefield. The most effective prompt description combination we have developed so far correctly classified 22 of the 52 gesture inputs provided (accuracy ≈ 40.74%), which is a significant improvement from our previous best approach correctly identifying 9 of the 52 gesture inputs (accuracy ≈ 17.31%). The main difference between those two approaches is that we added example descriptions, independently

written by each member, for all the gestures used in our gesture dataset. The main purpose for including these example descriptions is to provide the language models with a reference for the type of words and phrases we want the VLM output to be like and for the LLM classifier to classify against. Additionally, we provided the VLM with a system instruction to generate a frame-by-frame description, which led to descriptions being broken up into different movements over time. However, these differences are not enough to identify gestures consistently and correctly for use in the battlefield. We aim to push our gesture recognition framework to its limits by iterating and improving both our VLM and LLM prompts. Our current framework uses a zero-shot prompting approach, which means it does not rely on extensive training data and instead relies on well-designed, intentional prompts to perform well (Sahoo et al., 2024). We intend to guide our future iterations by conducting a literature review focusing on zero-shot and few-shot prompting strategies.

Once our gesture recognition framework consistently and correctly identifies gestures with accuracy we are satisfied with, our next steps will involve implementing the framework on a robot and expanding the framework to account for more variability. Our current gesture data is recorded under well-lit conditions, containing little visual noise or camera movement outside of the gesture, and contains only one person in the frame. For the gesture recognition pipeline to effectively work in the field, the robot must also have the ability to recognize gestures while the robot is moving in a visually noisier environment alongside multiple people, who may or may not be the person the robot is working alongside. We intend to explore methods and strategies to account for the increased variability through experimentation (trial and error) in addition to reviewing existing literature.

STS Discussion

The goal of my STS research is to develop a comprehensive analysis of the implications and ethical considerations of choice architecture in American healthcare from the perspectives of both patients and healthcare providers. I anticipate analyzing this from an actor network theory (ANT) framework, as both patients and providers are actors who are influenced by the options that are known/available to them and thus influence each other through their decision making. By focusing on the relationships between diverse types of choice architecture, patients, and providers from an ANT framework, the degree of influence each part has may be better understood in addition to identifying barriers and facilitators in the decision-making process of patients and providers. My primary research method will involve a literature review pertaining to articles about the decisions faced by Americans regarding decisions about insurance, which primary healthcare provider to choose, when to seek help regarding a sickness or problem, and other related choices. Additionally, my literature review will include articles about the decisions faced by healthcare professionals regarding how different treatment options are presented to patients, what information should be conveyed to other providers seeing the same patient, which types of insurance to accept, and other related choices. Two resources I have identified as worthwhile to understand in more depth are: *Does Dissatisfaction With Health Plans Stem From Having No Choices?* by Gawande et al. and *What Does the Evolution From Informed Consent to Shared Decision Making Teach Us About Authority in Health Care?* by Childress and Childress. While it is older, the first article contains information that is still relevant today about the impact of choice on a patient's satisfaction. The second article explores ethical considerations regarding informed consent, shared decision making, and patient autonomy which may influence what healthcare service providers consider when framing options to patients. The remainder of my

STS discussion will discuss my motivation for examining choice architecture from the perspective of both patients and providers in addition to key factors identified so far.

It is especially important to understand how different choices in the American healthcare system are presented to its constituents, as these can radically affect the level of care received and affect their perceived quality of care. One factor that greatly influences the options available to a patient is the socioeconomic status of a patient. Patients of a lower socioeconomic (SES) status have less access to and receive a lower quality of healthcare services compared to those of a higher socioeconomic status (Caballo et al., 2021). My literature review will primarily focus on the choice architecture faced by those of a lower SES because I believe my analysis will prove to be most beneficial for assisting those of a lower SES. Patients of a higher SES can more easily afford improved healthcare services, although they too are influenced by the choices known and available to them. Another factor that can influence the choice architecture from a patient's perspective is the patient's geographical location. While the United States has an immense amount of geographical variation, the most important distinction for the purposes of my analysis is whether a patient is from a rural or urban/suburban area. While patients of all geographical locations share common enablers and obstacles, there are some differences regarding care availability and ability to perceive care (Cyr et al., 2019). My literature review will account for both urban and rural environments equally, as while more people live in urban areas (meaning any meaningful change in the choice architecture will impact a greater number of patients), people living in rural areas are more disadvantaged from an accessibility standpoint and should not be left behind.

While gaining an effective understanding of a patient's choices is invaluable, it is just as necessary to comprehensively understand the perspective of the professionals providing

healthcare services. I will initially focus on evaluating how professionals currently make decisions regarding how to present different treatment options to patients in addition to how patient information is communicated to other healthcare institutions. Physicians and other health professionals have an ethical challenge regarding when to make decisions for the patient or share information and allow for patient autonomy (Childress & Childress, 2020). Patients may not always make decisions that are in their best interest from a health standpoint, leaving it up to the professional to consider each situation and decide how to frame information to a patient. This affects the type of care and treatments a patient will ultimately receive. Alongside providing information to just the patient, health professionals should be able to effectively communicate the patient's information to other institutions caring for that patient. Currently, care is poorly coordinated between these institutions for adults with serious illnesses or chronic conditions (Schoen et al., 2011). It would be useful to understand where breakdowns in coordination occur and investigate why the choice architecture faced by professionals does not allow or nudge for a potential solution to each breakdown cause, as less coordination breakdowns should lead to an improved quality of care for patients that need it the most.

Conclusion

In summary, our technical work aims to enable fluent human-robot collaboration through a gesture recognition framework which will allow a robot to work alongside a human by responding accordingly to different gestures. Our main obstacle is the prompt engineering associated with our zero-shot approach while next steps involve implementing the framework on a robot and accounting for greater variability in viewing conditions.

The primary focus for my STS research is to conduct a literature review analyzing the implications and ethical considerations of choice architecture in the American healthcare system

from the perspectives of both patients and healthcare providers. This analysis will be performed from an actor theory network framework as patients and providers alike are actors influenced by the options known to them. My analysis will consider the various backgrounds of patients such as socioeconomic status and geographical location alongside the ethical challenges faced by professionals providing care.

References

Benjdira, B., Koubaa, A., & Ali, A. M. (2023). *ROSGPT_Vision: Commanding Robots Using Only Language Models' Prompts* (No. arXiv:2308.11236). arXiv. http://arxiv.org/abs/2308.11236

Caballo, B., Dey, S., Prabhu, P., Seal, B., Chu, P., & Kim, L. (2021). *The Effects of Socioeconomic Status on the Quality and Accessibility of Healthcare Services*. https://doi.org/10.5281/ZENODO.4740684

Childress, J. F., & Childress, M. D. (2020). What Does the Evolution From Informed Consent to Shared Decision Making Teach Us About Authority in Health Care? *AMA Journal of Ethics*, *22*(5), 423–429. https://doi.org/10.1001/amajethics.2020.423

Cyr, M. E., Etchin, A. G., Guthrie, B. J., & Benneyan, J. C. (2019). Access to specialty healthcare in urban versus rural US populations: A systematic literature review. *BMC Health Services Research*, *19*(1), 974. https://doi.org/10.1186/s12913-019-4815-5

Ducharme, J. (2023, May 16). *Exclusive: More Than 70% of Americans Feel Failed by the Health Care System*. TIME. https://time.com/6279937/us-health-care-system-attitudes/

Fairchild, M. (2021, September 12). *Top Industries Using Robots | HowToRobot*. https://howtorobot.com/expert-insight/top-industries-using-robots

Gawande, A. A., Blendon, R. J., Brodie, M., Benson, J. M., Levitt, L., & Hugick, L. (1998). Does Dissatisfaction With Health Plans Stem From Having No Choices? *Health Affairs*, *17*(5), 184–194. https://doi.org/10.1377/hlthaff.17.5.184

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications* (No. arXiv:2402.07927). arXiv. https://doi.org/10.48550/arXiv.2402.07927

Schoen, C., Osborn, R., Squires, D., Doty, M., Pierson, R., & Applebaum, S. (2011). New 2011 Survey Of Patients With Complex Care Needs In Eleven Countries Finds That Care Is Often Poorly Coordinated. *Health Affairs*, *30*(12), 2437–2448. https://doi.org/10.1377/hlthaff.2011.0923

Schoen, C., Osborn, R., Squires, D., & Doty, M. M. (2013). Access, Affordability, And Insurance Complexity Are Often Worse In The United States Compared To Ten Other Countries. *Health Affairs*, *32*(12), 2205–2215. https://doi.org/10.1377/hlthaff.2013.0879

Shi, H., Ye, S., Fang, X., Jin, C., Isik, L., Kuo, Y.-L., & Shu, T. (2024). *MuMA-ToM: Multi-modal Multi-Agent Theory of Mind* (No. arXiv:2408.12574). arXiv. http://arxiv.org/abs/2408.12574