Towards a Common Metrics and Evaluation Framework for Assessment of Older Adults and Caregivers Interacting with Artificial Intelligence*

Jasmin Marwad[1†], Daisy M. Kiyemba[1†], Elizabeth J. Carter[2], Adam Norton[1]

[1]New England Robotics Validation and Experimentation (NERVE) Center,
University of Massachusetts Lowell, Lowell, Massachusetts, USA
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
jasmin_marwad@student.uml.edu, daisy_kiyemba@student.uml.edu, ejcarter@andrew.cmu.edu, adam_norton@uml.edu

Abstract

Artificial intelligence (AI) has applications in assisting older adults to age in place and provide support to them and their care givers as their cognition declines with age. However, effective assessment methods of this technology are needed in order to benchmark their performance and a common set of metrics and evaluation methods would enable such assessments to be compared to one another. To this end, we propose a common framework for human-AI interaction involving care recipients and their care networks. From the results of a literature review exercise, a framework with sample metrics, related measures, qualified evaluation tools, and contextual factors that impact assessment are reviewed. This paper provides a sample of common metrics in one of the framework's measurement spaces (human-AI interaction) and discusses some of the impacts of contextual factors and how use of the common metrics and evaluation framework can be used for meta-analysis and to guide future research. Additional future articles are planned to cover the other measurement spaces in the framework (system performance, task performance, and well-being), including their particular common metrics and evaluation methods. This effort aims to provide guidance for researchers in this domain as well as highlight measurement gaps that can be filled by future research.

Introduction

Advanced technologies with artificial intelligence (AI) have demonstrated their potential at enhancing the lives of older adults to allow them to age in place and support their quality of life despite cognitive decline. In addition to the care recipients, considerations for how these technologies can impact the caregivers (e.g., doctors, nurses, family members, friends) and the larger collaborative network of people must be taken. Effective measurement techniques are needed in order to quantitatively (and qualitatively) understand the progress being made, of which there are many in this domain and those adjacent to it. Additionally, commonality in metrics and evaluation methods is needed to: (1) allow for comparisons across studies, (2) provide guidance to new researchers, (3) encourage researchers to utilize validated assessment tools and dissuade them from developing novel ones, and (4) highlight gaps in measurement science to chart new research areas for the future.

Given the intersection of fields for the use case – such as human-computer interaction (HCI), human-robot interaction (HRI), medicine, computer science, and human factors – there are already many examples of performance evaluation that can be culled from. To this end, we propose a common metrics and evaluation framework for the assessment of older adults (care recipients) and their caregivers when interacting with AI. This paper proposes a series of categories and a sample set of metrics for human-AI interaction (HAI). This work is still under active development and additional articles covering the entire proposed metrics and evaluation framework are planned for the future.

Broader Impact

The metrics and evaluation framework is used to guide evaluations for the research conducted by the AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups (AI-CARING), whose mission is to develop the next generation of personalized collaborative AI systems that improve the quality of life and independence of aging adults living at home (AI-CARING 2024). At the time of publication, research within the institute is largely undergoing usability testing and deployment with older adults and their caregivers. Adoption of the framework is underway, but continued exercising of it and the common metrics is needed to refine it. The literature review that informed much of the framework focused on research that involved older adults of varying capabilities including mild cognitive impairment (MCI) and dementia. As AI is developed to improve the ability for people to age in place, effective and understandable assessment of each technology and experiment is needed in order to characterize its positive and negative impacts. The proposed framework aims to fill this gap by uniting the various areas for evaluation in the target application, providing a common lexicon for performance evaluation across the multiple research domains it encompasses.

*This work was supported by the National Science Foundation (IIS2112633).

†These authors contributed equally.

Related Work

There are many prior examples of metrics and evaluation frameworks used across the various fields that this particular application involves. For example, specific to AI in healthcare settings, (Reddy et al. 2021) uses capability, utility, and adoption as the three main components of evaluating the integration of AI into these applications. (Tanguay et al. 2023) is an example of an application-specific framework that focuses on the performance of AI in radiology, evaluating factors like quality, efficiency, and costs, with several software-specific metrics including accuracy and reliability. Others in the clinical domain focus on the subjective evaluation of the users, including ease of use, acceptance, satisfaction, and perceived benefit of use (Ji et al. 2021). The development and evaluation of explainable AI (XAI) systems in multiple contexts have produced assessment frameworks including the model usability evaluation (MUsE) framework (Dieber and Kirrane 2022) and XAI metrics that outline the measurable qualities of an explanation including goodness, satisfaction, and validity (Hoffman et al. 2018).

For embodied AI systems like robots, the foundational work in (Steinfeld et al. 2006) produced a common set of HRI metrics that relate to the performance of the robot, the impact of the communication factors, as well as task and operator performance. Building off of this broad framework, more specific efforts have been undertaken that include the definition of metrics to evaluate the effectiveness of interfaces in collaborative manufacturing (Marvel et al. 2020) and assessing the effectiveness of robot proficiency self-assessment and communication of proficiency (Norton et al. 2022). The latter provides a framework designed around each stage of interaction with a proficiency-based system, specifying metrics at each stage, thus uniting aspects of system performance and HRI.

The metrics and evaluation framework proposed in this paper aims to unite multiple measurement spaces relevant to the application and leverages similar structures to many of these related works.

Methodology

We conducted a semi-structured literature review of 350+ papers culled from the fields of medicine, psychology, computer science, robotics, and human factors by searching multiple online databases (e.g., Google Scholar, ACM Digital Library, IEEE Xplore, etc.) with a loose set of search terms (e.g., older adults, AI systems, etc.) to identify publications that covered studies and reviews. Initially, this review was limited to publication dates within 2000 to 2024, but that range was expanded as the citations within review papers were researched as well as incorporating external commendations from colleagues. This flexible approach was adopted as we intend to continue this literature review exercise as more work as published.

The reviewed literature was cataloged in a database according to various aspects of the studies conducted, including the type of AI technology used (if any), participant population, tasks performed during experimentation, and other contextual information. We have analyzed this ever-growing database to produce a dataset of qualified evaluation tools, distilled common metrics specifications, and the proposed framework that unites these assets. Readers are strongly encouraged to review the studies that are cited throughout this paper for more detail on the various metrics, measures, and tools that are discussed.

Tier	**Rec**	**Cronbach**	**Used with older adults**	**Used in experiment settings**
1	High	$\alpha \geq 0.7$	With MCI	Interacting with AI
2	Med	$\alpha \geq 0.7$	With or without MCI	Interacting with or without AI
*	Low	None reported		
3	Very low	$\alpha < 0.7$		

Table 1: Tier definitions and levels of recommendation used in the evaluation tools dataset from (Kiyemba et al. 2024).

The **evaluation tools dataset** (Kiyemba et al. 2024) consists of 240+ tools categorized as being used to assess either: cognitive ability; demographics, personality, and experiences; activity level; state of mind; or perceptions of the AI; and noting if training is required to use the tool. Individual instances (560+) of the tools being used in studies were then qualified based on whether or not the study involved interaction with AI/technology (i.e., if the participants in the study utilized an AI system to assist them in performing one or more tasks), if the participants included older adults (with MCI, dementia, or not cognitive impairment), and/or their caregivers, and any reported internal reliability scores. The dataset also indicates if a modified version of the original tool was used. Qualifications are distilled into tiers (1, 2, *, or 3) which corresponds to their applicability to use with older adults with MCI interacting with AI systems. Tier definitions can be seen in Table 1), with each corresponding to a level of recommendation:

- Tier 1: high; matches all relevant criteria for reliability and domain.
- Tier 2: medium; further research is needed for tool validation through experimentation with AI systems and/or the target population.
- Tier *: low; further research is needed for tool validation and reliability.
- Tier 3: very low; further research is needed not only for tool validation and reliability, but also validation when using AI systems and/or the target population.

After conducting the literature review and extracting all metrics evaluated per study, the metrics were analyzed and distilled in order to merge metrics that evaluated similar aspects of an interaction while including various examples of measures and tools used that evaluated the metric differently. This exercise was an attempt to limit the number of unique metrics named for simplicity of identifying commonalities while also allowing flexibility in how metrics are evaluated

given the variety of research goals and AI capabilities deployed during experimentation. This information was also used to determine the four unique measurement spaces used in the framework, covered in the next section.

Framework

The depth of each of the metrics and evaluation efforts described in the "Related Work" section varies widely, with some covering the performance of the AI system, the human's experience interacting with the system, and/or the tasks they perform together. The framework this paper proposes unites each of these areas with particular focus on the application to HAI involving older adults and caregivers, adding an axis for how the AI impacts the human beyond the interaction.

The proposed metrics and evaluation framework (see Figure 1) is divided into four **measurement spaces** that classify the type of assessment being conducted:

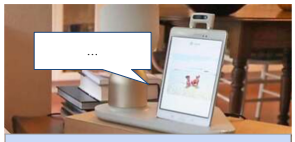
- Human-AI Interaction (HAI): evaluating the interactions between one or more humans (e.g., person with MCI, professional care provider, informal care provider, other) and the developed AI technology.
- System Performance (SP): evaluating the performance of the developed AI technology either as part of a human-AI interaction or a separate experiment (e.g., using an existing dataset).
- Task Performance (TP): evaluating the performance of tasks by one or more humans, systems, or both as part of an experiment including activities of daily living (ADLs).
- Well-Being (WB): evaluating a user's state of comfort, health, or happiness, either self-reported by the user or reported by a third party (such as their caregiver) with their interpretation of the user's well-being.

Within each space, a series of metrics are specified with associated tools (if applicable/available) and measures that can be used to evaluate that metric. A series of **contextual factors** are outlined that should be considered when selecting appropriate measures and tools as they impact how evaluations are conducted, including:

- Technology: augmented reality, computer, phone/tablet, robot, simulation/videos, wearable, virtual reality.
- Population: age, gender, ethnicity, occupation.
- Stakeholders involved: care recipient, professional caregiver, informal caregiver.
- Cognitive ability of participants: (if care recipients) MCI, dementia, Alzheimer's, Parkinson's, traumatic brain injury, physical disabilities, no impairment.
- Application tasks: basic, instrumental, or enhanced ADLs, or other task.
- Setting/environment: hospital, nursing home, research lab, home, simulation, remote/online, public space.
- Timeframe: number, frequency, and length of sessions.
- Methods: survey, interview, observation, experiment.

For this domain, the evaluation of **activities of daily living (ADLs)** is incredibly important as they are typically how

Human-AI Interaction (HAI) Metrics	
Acceptance	Attitudes
Engagement	Level of Interaction
Perception	Preference
Satisfaction	Self-Efficacy
Trust	Usability



System Performance (SP) Metrics	
Accuracy	Efficiency
Resiliency	Responsiveness
Robustness	Speed
…	

Task Performance (TP) Metrics	
Accuracy	Efficiency
Level of Performance: Ability, Capability, Correctness, Difficulty, Function,	

Well-Being (WB) Metrics	
Activity Level	Emotional State
Level of Stress and Burden	Quality of Life

Contextual Factors			
Technology	Population	Stakeholders Involved	Cognitive Ability
Application Tasks	Setting/Environment	Timeframe	Methods

Figure 1: The proposed metrics and evaluation framework.

the performance of tasks and degradation thereof is tracked as older adults age and potentially develop cognitive impairments. The three categories and their corresponding ADLs are as follows:

- Basic (B-): Ambulating (B-A), Feeding (B-F), Dressing (B-D), Personal hygiene (B-P), Continence (B-C), Toileting (B-T)
- Instrumental (I-): Transportation and shopping (I-TS), Managing finances (I-MF), Meal preparation (I-MP), Housecleaning and home maintenance (I-HM), Managing communication with others (I-MC), Managing medications (I-MM)
- Enhanced (E-): Hobbies (E-H), New learning (E-NL), Social communication (E-S)

Each of the contextual factors can affect how an evaluation should be conducted, which metrics and measures to use, what tools will be most effective, etc. For example, when evaluating WB: Level of Stress and Burden, if the stakeholders involved (i.e., the participants in the study being evaluated) are care recipients, then the State-Trait Anxiety Inventory (STAI) can be used (tier *) (Pino et al. 2020), but if caregivers are being evaluated, then the Caregiver Burden Scale (CBS) may be more appropriate (tier 2) (Fuh et al. 1999). Depending on the application task(s) selected for an experiment, each ADL has different performance criteria that can be evaluated. For example, the Assessment of Motor and Process Skills (AMPS) method provides criteria for performing 125 standardized ADL tasks (Center for Innovative OT Solutions 2023), which can be used to measure TP: Accuracy, Efficiency, and Level of Performance.

Sample of Common Metrics

This paper does not cover all proposed common metrics; rather, summaries of the proposed common metrics only in the HAI measurement space are provided as a sample. Future articles are under development that will provide more thorough reviews of common metrics within the framework for HAI, TP, and WB. Aside from a general outline of common SP metrics, we do not intend to cover all possible measures and evaluation tools (e.g., published benchmarks for comparison) for SP due to high variety of AI technologies that are applicable to this use case, each with their own considerations. For each metric reviewed in this section, a brief definition is provided, noting if the associated measures are subjective and/or objective, can be evaluated a priori or post hoc, with references to example studies that evaluated one or more ADLs using the qualified evaluation tools from our dataset. A summary of the sample HAI metrics, tools, and measures reviewed can be found in Table 2.

Acceptance measures a user's ability to accept and use the AI technology. Measures of acceptance are largely subjective and administered post hoc wherein the participant is asked to provide an acceptance rating towards an AI/technology. There are several examples in the literature for subjective evaluation of acceptance post hoc for managing communication with others (I-MC) (Hossain and Ahmed 2012; Luperto et al. 2022), managing medications (I-MM) (Luperto et al. 2022), and feeding (B-F) (McColl and Nejat 2013; Di Napoli, Ercolano, and Rossi 2022), using evaluation tools including the Almere Model (tier 1), the Unified Theory of Acceptance & Use of Technology (UTAUT; tier 1), and the Robot Acceptance Questionnaire (tier 2).

Attitudes assesses a participant's overall feelings, beliefs, reactions, and perspectives towards the AI technology. This is exclusively a subjective measure and can be evaluated both a priori and post hoc, often comparing the two. Examples in the literature that utilized ADLs include (Mitzner et al. 2010) which evaluated all IADLs and (Stafford et al. 2014) for managing communication with others (I-MC) only. The latter utilized the Robot Attitudes Scale (RAS; tier 2). Several other tools are available including the Negative Attitudes towards Robots Scale (NARS; tier 1) (Johansson-Pajala et al. 2022), the Computer Attitudes Scale (tier 2) (Mitzner et al. 2019), and the Positive Attitudes Towards Robots (tier 2) (Rantanen et al. 2020) which is an inverse of the NARS.

Engagement measures the extent of the participants' active or passive involvement with the technology while it is being used. This metric can be measured subjectively via established tools or objectively by observing physical actions. For subjective measures, (Tulsulkar et al. 2021) used tools such as the Observed Emotion Rating Scale (OERS) and Menorah Park Engagement Scale (MPES) (both tier *) during social communication (E-SC). The same study also used The Almere Model (tier *) for post hoc subjective evaluation. Examples of objective measures include those to evaluate E-SC such as the number and length of conversations between the participant and the AI/technology (Abdollahi et al. 2017, 2022; Fan et al. 2021; McColl and Nejat 2013) or the number of times a caregiver extended their hand towards the robot (Kramer, Friedmann, and Bernstein 2009).

Level of Interaction assesses the degree to which a participant actively interacts with an AI/technology during an experiment. Subjective evaluation use tools like UTAUT (tier 1) during social communication (E-SC) (Heerink et al. 2010) or the Borg Rating of Perceived Exertion (RPE; tier 3) when performing hobbies (E-H) (Fitter et al. 2020). Objective measures include more detailed assessments of conversational speech analysis to measure characteristics like length of utterances, silence, filler, jitter, and shimmer during E-SC (Yoshii et al. 2021) or measuring the number of times participants used a robot to assist with managing communications with others (I-MC) (Stafford et al. 2014). Level of Interaction and Engagement were both the second most evaluated HAI metric in the studies reviewed.

Perception measures components of how users think about and interpret the behavior of AI/technology and its impact on daily activities. All examples in the literature review were subjectively evaluated post hoc, such as in (Mois and Beer 2020) wherein the Robotic Social Attributes Scale (RoSAS) and the Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology tool (both tier *) were used to assess performance of hobbies (E-H). Another study used the Dimensions of Mind Perception Scale (tier 3) to evaluate a service robot supporting execution of managing medications (I-MM), managing communications with others (I-MC), and new learning (E-NL) (Stafford et al. 2014).

Preference refers to the user's subjective selection and favor of one option over another (e.g., AI/technology system, feature, or method) based on its ability to meet their specific needs, ease of use, and/or the overall enhancement it provides. Amongst the HAI metrics used in studies found during the literature review, Preference was most frequently evaluated. For example, evaluating user preference of how a robot delivers their medications for managing medications (I-MM) (Prakash et al. 2013), assists with house cleaning for housecleaning and home maintenance (I-HM) (Beer et al. 2017), or voice-based interface design choices for AI system assisting with transportation and shopping (I-TS) and manging communication with others (I-MC) (Granata et al. 2010). Some of the most commonly used tools include the Robot Opinions Questionnaire and the Assistance Preference Checklist (both tier 3), both of which were used to evaluate preferences of older adults on robots assisting with all ADLs in (Smarr et al. 2012, 2014).

Satisfaction is measured to determine the extent to which users feel the AI/technology meets their needs, evaluated subjectively as a post hoc measure after interacting with the AI/technology. In (Viswanathan et al. 2011), the Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST 2.0; tier *) was used to evaluate satisfaction with using robotic wheelchair for navigation (i.e., ambulating [B-A]), while other studies use unvalidated scales such as a Likert scale ranging from "not at all" to "very much" (Hughes et al. 2014) after playing an interactive video game (i.e., hobbies [E-H]).

Self-Efficacy measures a user's self-assessed evaluation of their capacity and ability to complete certain tasks, pos-

Metric	Type	Associated tools and/or measures	Example study citations	Tier	ADLs
Acceptance	Sub	The Almere Model	(McColl and Nejat 2013)	1	B-F
		Unified Theory of Acceptance & Use of Technology (UTAUT)	(Di Napoli, Ercolano, and Rossi 2022)	1	I-MM
		Robot Acceptance Questionnaire	(McColl and Nejat 2013)	2	B-F
		Rating acceptance 1 to 7	(Hossain and Ahmed 2012)	n/a	I-MC
		Open responses during conversation and monitoring social interactions	(Luperto et al. 2022)	n/a	
			(Hebesberger et al. 2017)	n/a	E-SC
Attitudes	Sub	Robot Attitudes Scale (RAS)	(Stafford et al. 2014)	2	I-MC
		Positive Attitudes Towards Robots	(Rantanen et al. 2020)	2	n/a
		Computer Attitudes Scale	(Mitzner et al. 2019)	2	n/a
		Negative Attitudes towards Robots Scale (NARS)	(Johansson-Pajala et al. 2022)	1	n/a
Engagement	Sub	The Almere Model	(Moro et al. 2019)	*	I-MP
		Observed Emotion Rating Scale (OERS); Menorah Park Engagement Scale (MPES)	(Tulsulkar et al. 2021)	*	E-SC
	Obj	Number and length of conversations with AI/technology	(Abdollahi et al. 2017, 2022; Fan et al. 2021)	n/a	E-SC
		Participant compliance with robot reminders and prompts	(Søraa et al. 2021)	n/a	I-HM, I-MM
			(Begum et al. 2013)	n/a	I-MP
Level of Interaction	Sub	Unified Theory of Acceptance & Use of Technology (UTAUT)	(Heerink et al. 2010)	1	E-SC
		Borg Rating of Perceived Exertion (RPE)	(Fitter et al. 2020)	3	E-H
	Obj	Number of robot usages	(Stafford et al. 2014)	n/a	I-MC
		Analyzing conversational speech for utterance length, filler, jitter, etc.	(Yoshii et al. 2021)	n/a	E-SC
Perception	Sub	Robotic Social Attributes Scale (RoSAS)	(Mois and Beer 2020)	*	E-H
		Dimensions of Mind Perception Scale	(Stafford et al. 2014)	3	I-MC, I-MM, E-NL
Preference	Sub	Robot Opinions Questionnaire; Assistance Preference Checklist	(Smarr et al. 2012, 2014)	3	All ADLs
			(Prakash et al. 2013; Beer et al. 2017)	3	I-MM
Satisfaction	Sub	From "not at all" to "very much"	(Hughes et al. 2014)	n/a	E-H
		Quebec User Evaluation of Satisfaction with Assistive Technology (QUEST 2.0)	(Viswanathan et al. 2011)	*	B-A
Self-Efficacy	Sub	Daily Living Self-Efficacy Scale (DLSES)	(Stuck and Rogers 2017, 2019)	2	B-P, I-HH, I-MF, I-TS, E-H
		Robot Usage Self-Efficacy; Robot Familiarity and Use Questionnaire	(Stuck and Rogers 2017)	3	
Trust	Sub	Trust in Medical Technology Scale	(Mann et al. 2015)	2	n/a
		Trust in Automation (TIA) Questionnaire	(Körber 2019)	2	n/a
		Trust in Assistance Checklist; Propensity to Trust Scale	(Langer et al. 2019)	*	n/a
			(Stuck and Rogers 2017)	3	B-D, B-P, I-HM, I-MM, I-MP
Usability	Sub	System Usability Scale (SUS)	(Barg-Walkow et al. 2017)	2	E-H
		Unified Theory of Acceptance & Use of Technology (UTAUT)	(Heerink et al. 2010)	1	n/a
		Open response feedback on usability	(Wu et al. 2017; Seelye et al. 2012)	n/a	I-MC
	Obj	Task success rate: accepting a video call, mute and unmute the microphone and speakers, etc.			

Table 2: Summary of a subset of the sample common metrics for HAI.

sess knowledge of specific skills, and the ability to carry them out. This is also often referred to as "confidence" in the literature. For example, in (Stuck and Rogers 2017) older adults' self-efficacy was evaluated when performing task related to personal hygiene (B-P), managing finances (I-MF), transportation and shopping (I-TS), housecleaning and home maintenance (I-HM), and hobbies (E-H), using the Robot Usage Self-Efficacy tool (tier 3), the Robot Familiarity and Use Questionnaire (tier 3), and the Daily Living Self-Efficacy Scale (DLSES; tier 2). The latter tool was also used to evaluate the same ADLs in a follow-on study (Stuck and Rogers 2019).

Trust, while its own research field, there have been studies that attempt to utilize established trust scales for the target domain. For example, (Stuck and Rogers 2017) utilized the Trust in Assistance Checklist and the Propensity to Trust Scale (both tier 3) to evaluate whether older adults would prefer to trust a human or robot to assist with dressing (B-D), personal hygiene (B-P), housecleaning and home maintenance (I-HM), managing medications (I-MM), and meal preparation (I-MP). The same two tools were also used in (Langer et al. 2019) for using robots in rehabilitation (tier *). There are many other trust tools available with varying applicability to this domain, including the Trust in Medical Technology Scale (tier 2) (Mann et al. 2015) and the Trust in Automation (TIA) questionnaire (tier 2) (Körber 2019).

Usability measures the degree to which an AI/technology can be used effectively and easily, the latter of which is often referred to as "ease of use." Usability can be evaluated subjectively and objectively. For subjective evaluation, tools like Unified Theory of Acceptance & Use of Technology (UTAUT) (tier 1) (Heerink et al. 2010) or the System Usability Scale (SUS) to evaluate performing hobbies (E-H) (tier 2) (Barg-Walkow et al. 2017) are used. Objective measures typically refer to the successful completion of tasks, such as participants answering and making calls, powering the device on and off, or adjusting its settings, performing tasks related to managing communication with others (I-MC) (Seelye et al. 2012).

Discussion

Using this framework, the results of one or more studies can be compared, with alignments and differences in metrics, measures, and context described using a common lexicon. Effectively all experimental studies conducted in this domain (HAI with older adults and caregivers) can be characterized using the framework, enabling meta-analyses to be conducted of the state-of-the-art. For example, of the literature review conducted, it was found that the three most common HAI metrics evaluated were Engagement, Level of Interaction, and Preference. When investigating studies that evaluated Engagement, such as (Abdollahi et al. 2017) and (Søraa et al. 2021), both used objective measures of this metric, but in different contexts (humanoid robot in a nursing home in the United States vs. flower pot robot in residential homes in the Netherlands, Italy, and Switzerland, respectively). The goals of each study also differed, with the former relying largely on HAI evaluation (linking Engagement to Acceptance of life-like robots) while the latter also included WB metrics (evaluating domestication of technology and its impact on social relations, linking HAI: Engagement to WB: Quality of Life). This also demonstrates how the relationships between each metrics within each measurement space or across multiple spaces can be assessed. When conducting evaluations during development, correlations found between metrics (e.g., HAI: Engagement of certain types of care recipients – as characterized by the contextual factors axes – and the SP: Efficiency of the AI system) may result in system design revisions to improve its effectiveness.

Some preliminary takeaways of the HAI measurement space have been revealed thus far for future investigation. Given the number of subjective metrics and evaluation tools used, and the target population's potential for limited interaction means and understanding (e.g., due to cognitive decline or impairment), it is important to note what target stakeholder is under evaluation compared to the stakeholder who is physically responding or inputting the response. For example, if the target stakeholder for an AI technology is a care recipient, it is not uncommon for a caregiver to participate alongside them helping them to understand the questions being asked and respond accurately. In other cases, a caregiver may respond independently of the care recipient and provide their own analysis of the care recipient. A common language must be developed in order to categorize and understand the context of how these subjective metrics are derived.

Another takeaway is that, in general, evaluation of the caregiver experience while interacting with AI is significantly under-researched compared to that of the care receiver. There are some evaluation tools for this type of assessment, though, including the Caregiver Burden Scale (CBS) qualified as tier 2 in (Fuh et al. 1999) and the Zarit Burden Interview (ZBI) qualified as tier * in (Inoue, Wada, and Shibata 2021). Given the integral role that caregivers play in a care recipient's quality of life and the larger network of caregivers, their experiences must be properly evaluated in order to ensure AI/technology for this domain will be accepted and effective.

Conclusion

This paper proposes a metrics and evaluation framework for assessing older adults (i.e., care recipients) and caregivers interacting with AI. A sample of metrics from one of the framework's measurement spaces (HAI) are covered briefly with several examples of measures and evaluation tools used in the literature. Many more metrics remain to be analyzed in the other spaces (SP, TP, and WB) in future publications. Some examples of the impact of various contextual factors are provided, but many more are still to be explored. By proliferating this framework throughout the research domain, dedicated studies can be run that investigate these impacts. Through this exercise, we aim to further identify the trends, strengths, and limitations of previously used measurement techniques towards the development of common metrics and evaluation methods to assess older adult care recipients and their caregivers interacting with AI.

References

Abdollahi, H.; Mahoor, M.; Zandie, R.; Sewierski, J.; and Qualls, S. 2022. Artificial emotional intelligence in socially assistive robots for older adults: a pilot study. *IEEE Transactions on Affective Computing*.

Abdollahi, H.; Mollahosseini, A.; Lane, J. T.; and Mahoor, M. H. 2017. A pilot study on using an intelligent life-like robot as a companion for elderly individuals with dementia and depression. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 541–546. IEEE.

AI-CARING. 2024. AI Institute for Collaborative Assistance and Responsive Interaction for Networked Groups (AI-CARING). https://ai-caring.org/. Accessed: 2024-08-27.

Barg-Walkow, L. H.; Harrington, C. N.; Mitzner, T. L.; Hartley, J. Q.; and Rogers, W. A. 2017. Understanding older adults' perceptions of and attitudes towards exergames. *Gerontechnology: international journal on the fundamental aspects of technology to serve the ageing society*, 16(2): 81.

Beer, J. M.; Prakash, A.; Smarr, C.-A.; Chen, T. L.; Hawkins, K.; Nguyen, H.; Deyle, T.; Mitzner, T. L.; Kemp, C. C.; and Rogers, W. A. 2017. Older users' acceptance of an assistive robot: Attitudinal changes following brief exposure. *Gerontechnology: international journal on the fundamental aspects of technology to serve the ageing society*, 16(1): 21.

Begum, M.; Wang, R.; Huq, R.; and Mihailidis, A. 2013. Performance of daily activities by older adults with dementia: The role of an assistive robot. In *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*, 1–8. IEEE.

Center for Innovative OT Solutions. 2023. AMPS Task Notes Generator. https://www.innovativeotsolutions.com/tools/amps/tasks/. Accessed: 2023-06-12.

Di Napoli, C.; Ercolano, G.; and Rossi, S. 2022. Personalized home-care support for the elderly: a field experience with a social robot at home. *User Modeling and User-Adapted Interaction*, 1–36.

Dieber, J.; and Kirrane, S. 2022. A novel model usability evaluation framework (MUsE) for explainable artificial intelligence. *Information Fusion*, 81: 143–153.

Fan, J.; Mion, L. C.; Beuscher, L.; Ullal, A.; Newhouse, P. A.; and Sarkar, N. 2021. SAR-connect: a socially assistive robotic system to support activity and social engagement of older adults. *IEEE Transactions on Robotics*, 38(2): 1250–1269.

Fitter, N. T.; Mohan, M.; Kuchenbecker, K. J.; and Johnson, M. J. 2020. Exercising with Baxter: preliminary support for assistive social-physical human-robot interaction. *Journal of neuroengineering and rehabilitation*, 17: 1–22.

Fuh, J.-L.; Wang, S.-J.; Liu, H.-C.; and Wang, H.-C. 1999. The caregiving burden scale among Chinese caregivers of Alzheimer patients. *Dementia and Geriatric Cognitive Disorders*, 10(3): 186–191.

Granata, C.; Chetouani, M.; Tapus, A.; Bidaud, P.; and Dupourqué, V. 2010. Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders. In *19th International Symposium in Robot and Human Interactive Communication*, 785–790. IEEE.

Hebesberger, D.; Koertner, T.; Gisinger, C.; and Pripfl, J. 2017. A long-term autonomous robot at a care hospital: A mixed methods study on social acceptance and experiences of staff and older adults. *International Journal of Social Robotics*, 9(3): 417–429.

Heerink, M.; et al. 2010. Assessing Acceptance of Assistive Social Agent Technology by Older Adults: the Almere.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Hossain, M. A.; and Ahmed, D. T. 2012. Virtual caregiver: an ambient-aware elderly monitoring system. *IEEE Transactions on Information Technology in Biomedicine*, 16(6): 1024–1031.

Hughes, T. F.; Flatt, J. D.; Fu, B.; Butters, M. A.; Chang, C.-C. H.; and Ganguli, M. 2014. Interactive video gaming compared with health education in older adults with mild cognitive impairment: a feasibility study. *International journal of geriatric psychiatry*, 29(9): 890–898.

Inoue, K.; Wada, K.; and Shibata, T. 2021. Exploring the applicability of the robotic seal PARO to support caring for older persons with dementia within the home context. *Palliative Care and Social Practice*, 15: 26323524211030285.

Ji, M.; Genchev, G. Z.; Huang, H.; Xu, T.; Lu, H.; and Yu, G. 2021. Evaluation framework for successful artificial intelligence–enabled clinical decision support systems: mixed methods study. *Journal of medical Internet research*, 23(6): e25929.

Johansson-Pajala, R.-M.; Zander, V.; Gustafsson, C.; and Gusdal, A. 2022. No thank you to humanized robots: attitudes to care robots in elder care services. *Home Health Care Services Quarterly*, 41(1): 40–53.

Kiyemba, D. M.; Marwad, J.; Carter, E. J.; and Norton, A. 2024. Evaluation Tools for Human-AI Interactions Involving Older Adults with Mild Cognitive Impairments. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 915–918.

Körber, M. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, 13–30. Springer.

Kramer, S. C.; Friedmann, E.; and Bernstein, P. L. 2009. Comparison of the effect of human interaction, animal-assisted therapy, and AIBO-assisted therapy on long-term care residents with dementia. *Anthrozoös*, 22(1): 43–57.

Langer, A.; Feingold-Polak, R.; Mueller, O.; Kellmeyer, P.; and Levy-Tzedek, S. 2019. Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience & Biobehavioral Reviews*, 104: 231–239.

Luperto, M.; Monroy, J.; Renoux, J.; Lunardini, F.; Basilico, N.; Bulgheroni, M.; Cangelosi, A.; Cesari, M.; Cid, M.; Ianes, A.; et al. 2022. Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: the MoveCare project. *International Journal of Social Robotics*, 1–29.

Mann, J. A.; MacDonald, B. A.; Kuo, I.-H.; Li, X.; and Broadbent, E. 2015. People respond better to robots than computer tablets delivering healthcare instructions. *Computers in Human Behavior*, 43: 112–117.

Marvel, J. A.; Bagchi, S.; Zimmerman, M.; and Antonishek, B. 2020. Towards effective interface designs for collaborative HRI in manufacturing: Metrics and measures. *ACM Transactions on Human-Robot Interaction (THRI)*, 9(4): 1–55.

McColl, D.; and Nejat, G. 2013. Meal-time with a socially assistive robot and older adults at a long-term care facility. *Journal of Human-Robot Interaction*, 2(1): 152–171.

Mitzner, T. L.; Boron, J. B.; Fausset, C. B.; Adams, A. E.; Charness, N.; Czaja, S. J.; Dijkstra, K.; Fisk, A. D.; Rogers, W. A.; and Sharit, J. 2010. Older adults talk technology: Technology usage and attitudes. *Computers in human behavior*, 26(6): 1710–1721.

Mitzner, T. L.; Savla, J.; Boot, W. R.; Sharit, J.; Charness, N.; Czaja, S. J.; and Rogers, W. A. 2019. Technology adoption by older adults: findings from the PRISM trial. *The Gerontologist*, 59(1): 34–44.

Mois, G.; and Beer, J. M. 2020. The role of healthcare robotics in providing support to older adults: a socioecological perspective. *Current Geriatrics Reports*, 9: 82–89.

Moro, C.; Lin, S.; Nejat, G.; and Mihailidis, A. 2019. Social robots and seniors: A comparative study on the influence of dynamic social features on human–robot interaction. *International Journal of Social Robotics*, 11: 5–24.

Norton, A.; Admoni, H.; Crandall, J.; Fitzgerald, T.; Gautam, A.; Goodrich, M.; Saretsky, A.; Scheutz, M.; Simmons, R.; Steinfeld, A.; et al. 2022. Metrics for robot proficiency self-assessment and communication of proficiency in human-robot teams. *ACM Transactions on Human-Robot Interaction (THRI)*, 11(3): 1–38.

Pino, O.; Palestra, G.; Trevino, R.; and De Carolis, B. 2020. The humanoid robot NAO as trainer in a memory program for elderly people with mild cognitive impairment. *International Journal of Social Robotics*, 12: 21–33.

Prakash, A.; Beer, J. M.; Deyle, T.; Smarr, C.-A.; Chen, T. L.; Mitzner, T. L.; Kemp, C. C.; and Rogers, W. A. 2013. Older adults' medication management in the home: How can robots help? In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 283–290. IEEE.

Rantanen, T.; Leppälahti, T.; Porokuokka, J.; and Heikkinen, S. 2020. Impacts of a care robotics project on finnish home care workers' attitudes towards robots. *International Journal of Environmental Research and Public Health*, 17(19): 7176.

Reddy, S.; Rogers, W.; Makinen, V.-P.; Coiera, E.; Brown, P.; Wenzel, M.; Weicken, E.; Ansari, S.; Mathur, P.; Casey, A.; et al. 2021. Evaluation framework to guide implementation of AI systems into healthcare settings. *BMJ health & care informatics*, 28(1).

Seelye, A. M.; Wild, K. V.; Larimer, N.; Maxwell, S.; Kearns, P.; and Kaye, J. A. 2012. Reactions to a remote-controlled video-communication robot in seniors' homes: a pilot study of feasibility and acceptance. *Telemedicine and e-Health*, 18(10): 755–759.

Smarr, C.-A.; Mitzner, T. L.; Beer, J. M.; Prakash, A.; Chen, T. L.; Kemp, C. C.; and Rogers, W. A. 2014. Domestic robots for older adults: attitudes, preferences, and potential. *International journal of social robotics*, 6: 229–247.

Smarr, C.-A.; Prakash, A.; Beer, J. M.; Mitzner, T. L.; Kemp, C. C.; and Rogers, W. A. 2012. Older adults' preferences for and acceptance of robot assistance for everyday living tasks. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 56, 153–157. Sage Publications Sage CA: Los Angeles, CA.

Søraa, R. A.; Nyvoll, P.; Tøndel, G.; Fosch-Villaronga, E.; and Serrano, J. A. 2021. The social dimension of domesticating technology: Interactions between older adults, caregivers, and robots in the home. *Technological Forecasting and Social Change*, 167: 120678.

Stafford, R. Q.; MacDonald, B. A.; Jayawardena, C.; Wegner, D. M.; and Broadbent, E. 2014. Does the robot have a mind? Mind perception and attitudes towards robots predict use of an eldercare robot. *International journal of social robotics*, 6: 17–32.

Steinfeld, A.; Fong, T.; Kaber, D.; Lewis, M.; Scholtz, J.; Schultz, A.; and Goodrich, M. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 33–40.

Stuck, R. E.; and Rogers, W. A. 2017. Understanding older adult's perceptions of factors that support trust in human and robot care providers. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, 372–377.

Stuck, R. E.; and Rogers, W. A. 2019. Supporting trust in home healthcare providers: Insights into the care recipients' perspective. *Home Health Care Services Quarterly*, 38(2): 61–79.

Tanguay, W.; Acar, P.; Fine, B.; Abdolell, M.; Gong, B.; Cadrin-Chênevert, A.; Chartrand-Lefebvre, C.; Chalaoui, J.; Gorgos, A.; Chin, A. S.-L.; et al. 2023. Assessment of radiology artificial intelligence software: a validation and evaluation framework. *Canadian Association of Radiologists Journal*, 74(2): 326–333.

Tulsulkar, G.; Mishra, N.; Thalmann, N. M.; Lim, H. E.; Lee, M. P.; and Cheng, S. K. 2021. Can a humanoid social robot stimulate the interactivity of cognitively impaired elderly? A thorough study based on computer vision methods. *The Visual Computer*, 37: 3019–3038.

Viswanathan, P.; Little, J. J.; Mackworth, A. K.; and Mihailidis, A. 2011. Navigation and obstacle avoidance help (NOAH) for older adults with cognitive impairment: a pilot study. In *The proceedings of the 13th international*

ACM SIGACCESS conference on Computers and accessibility, 43–50.

Wu, X.; Thomas, R.; Drobina, E.; Mitzner, T.; and Beer, J. 2017. An evaluation of a telepresence robot: User testing among older adults with mobility impairment. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 325–326.

Yoshii, K.; Nishimura, M.; Kimura, D.; Kosugi, A.; Shinkawa, K.; Takase, T.; Kobayashi, M.; Yamada, Y.; Nemoto, M.; Watanabe, R.; et al. 2021. A study for detecting mild cognitive impairment by analyzing conversations with humanoid robots. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, 347–350. IEEE.