HRI Measures and Metrics

Prof. Wing Yue Geoffrey Louie and Suraj Goyal

Measures and Metrics

Although there is some overlap. What is the difference between a measure and metric?

Measure - Process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity**.** Must be observable, objective and concrete. This can include number of lines of code in software.

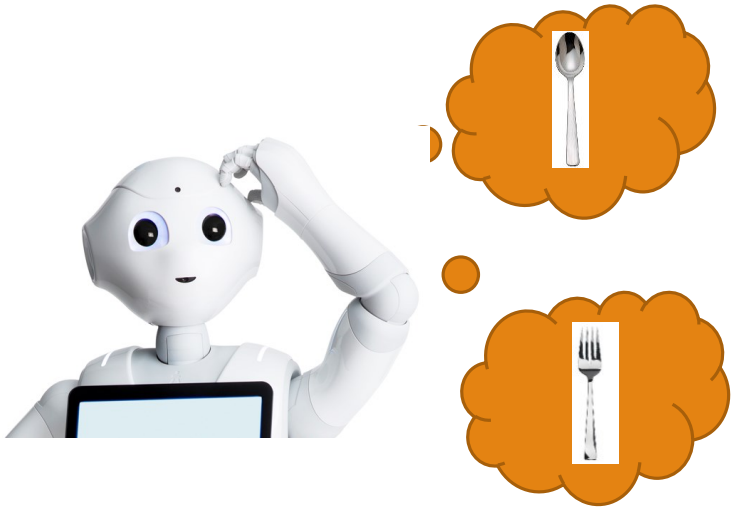
Metric – An abstract attribute to describe a system and require a series of measurements. Examples include robustness, reliability, efficiency, or effectiveness.

Why does this matter?

- Enables researchers to compare findings, benchmark designs, and have a common set of tools for evaluation

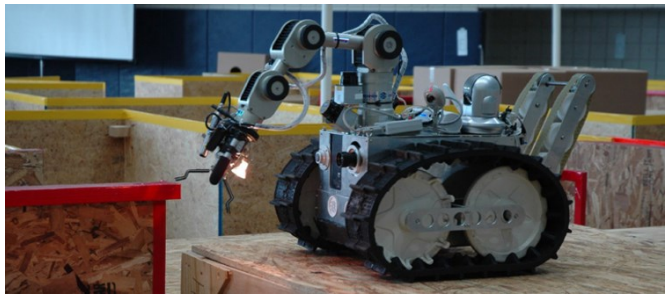
- Without a general set of metrics there is no focus for your design

- Enables you to optimize your design or create iterations to demonstrate progression



Challenges in Human-Robot Interaction

- Applications are numerous including: Assistive Robotics, USAR, Military, Space, Service
  - Metrics are not common across these applications
    - Time-to-completion is application or task specific so scenario-based or task-based reference tasks are used to evaluate designs
      - Example of this would be the Robocup Urban Search and Rescue Arena where you could test different control strategies or robot designs
        - Common measures can then be used for defining evaluating performance
          - # of victims found
      - Another common example is Blocks World or Towers Hanoi for high-level manipulation tasks
  - Even within the same application domain, tasks can have varying metrics.
    - What is considered effective for USAR? Completion time? Number of victims found? Areas searched?



Robocup Rescue

Robocup Home

Human-Robot Interaction Components for Evaluation

HRI can be evaluated from three different perspectives:

1) Human - In user-centered work, on the other hand, the focus of a study could be on understanding aspects of human behavior or cognition that will affect the success of HRI
2) Robot - In robot-centered work, the research focus might be on developing the technical capabilities that robots need to interact with people, or testing different aspects of the robot's functionality or design to see which are most effective
3) System – In system-centered work, we take a holistic approach considering the robot and human together in the human-robot interaction

We will discuss the:

1) Classes of metrics for comparison of research results
2) Identify common metrics used for evaluation across tasks and systems
3) Provide a general toolkit for studies

Common Human-Robot Interaction Tasks

The following are some common human-robot interaction tasks we will focus on:

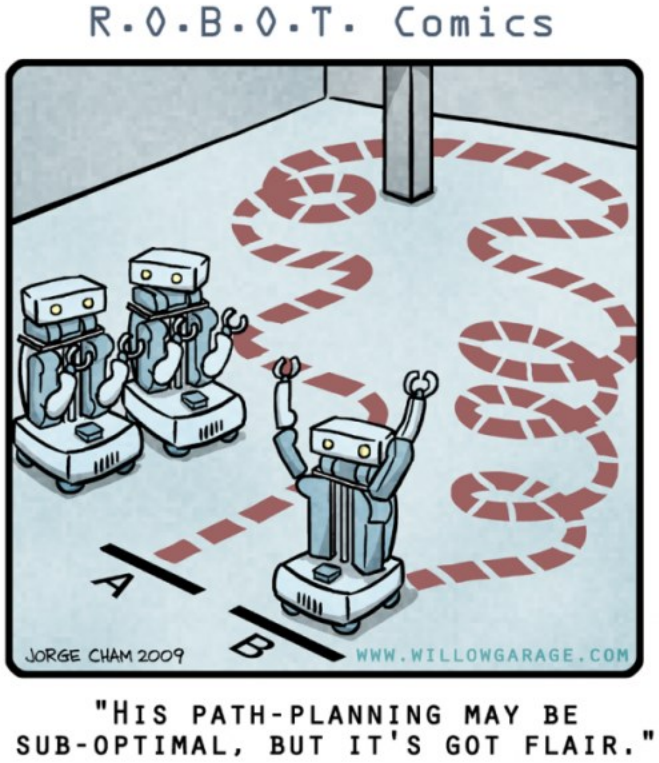
1) **Navigation**

Description of task: Moving from point A to point B

Components of task:

- Determining where robot is (localization)
- Determining where robot needs to go (goal)
- How robot should move (e.g. path, resource usage)
- How robot deals with environmental factors (obstacles, hazards)

Applications: Autonomous vehicles



Common Human-Robot Interaction Tasks

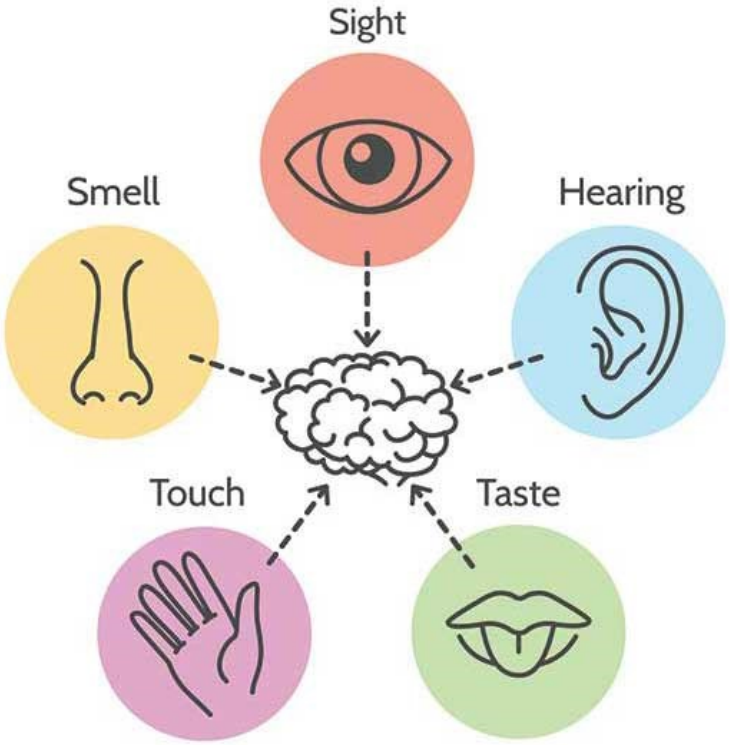
2) Perception

Description of task: Perceive and understand the environment such as search

Components of task:

- Establishing context through proprioceptive sensing
- Interpreting sensor data
- Seeking/filtering additional sensor data
- Deciding on information to provide other agents

Applications: Search, surveillance, target identification



Common Human-Robot Interaction Tasks

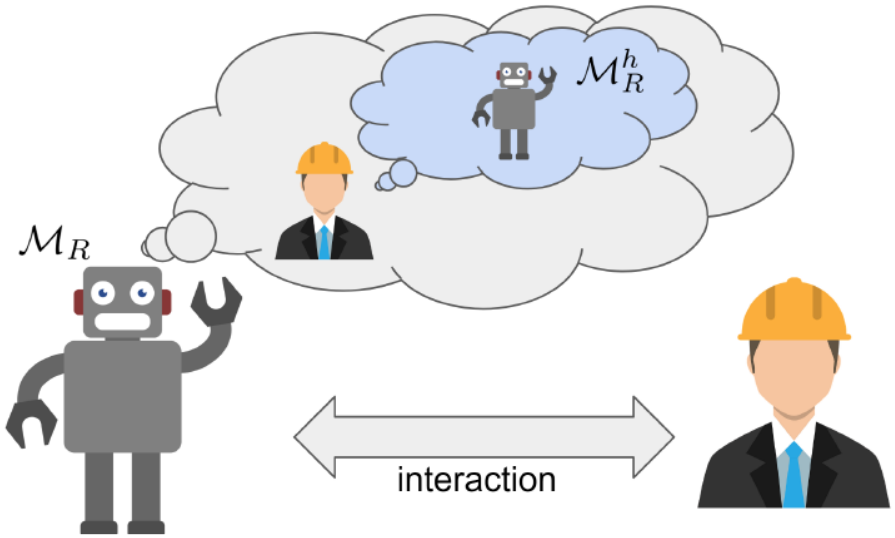
3) Management

Description of task: Coordinating and managing of actions for humans and robots to deploy and have resources at the right place at the right time.

Components of task:

- Assessing availability
- Understanding capabilities
- Team coordination
- Monitoring
- Recognizing problems
- Interventions

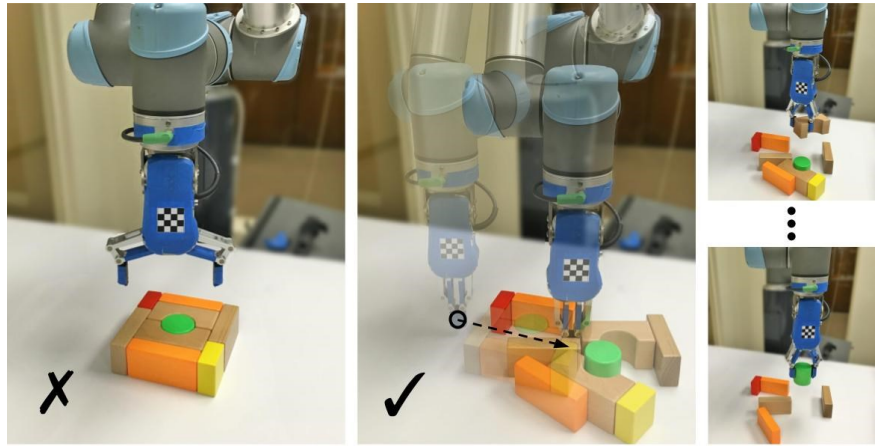
Applications: Human-robot teaming



Common Human-Robot Interaction Tasks

4) Manipulation

Description of task: Interacting with the environment to manipulate objects in the environment (e.g. grasp, push, release)



Components of task:

- Determining what is to be effected
- Specifying how it should be done
- Executing the process
- Verifying the outcome

Applications: Ordinance disposal, rock sampling, construction, material delivery

Common Human-Robot Interaction Tasks

5) Social

Description of task: Perform work that requires social interaction (e.g. verbal communication, non-verbal communication, proxemics)

Components of task:

- Perceiving and interpreting world based on past experiences
- Recognizing and modelling users
- Understanding social communication and norms
- Acquiring and exhibiting social competencies

Applications: Tour guide, healthcare, entertainment

Potential Biasing Effects

What is biasing effect?

“Bias may be defined as any systematic error that results in an incorrect estimate of the true effect of an exposure on the outcome of interest.”

Potential Sources of Biases in HRI

1) Communications – How information relayed between humans and robots

2) Robot Response – Variations in robot timings in response

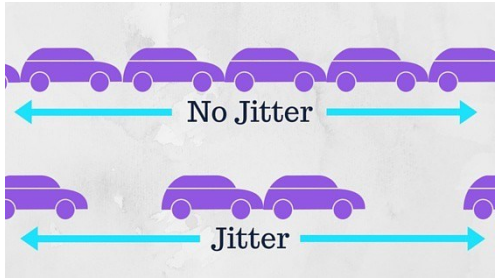
3) User – Factors which can influence behavior and affect human performance

Potential Sources of Biases in HRI - Communication

Communication factors which have confounding effects on human performance include:

Delay – This is a time delay for transmission of information due to communication networks. This is something commonly seen in space robotics and has commonly had negative effects on human performance.

Jitter – This refers to variations in time when a message is received. Creates a slingshot effect that humans have a difficult time adjusting to.



Bandwidth – The amount of data that can be transferred across a communication channel. This can result in loss of sensory details because of data compression measures

Potential Sources of Biases in HRI – Robot Response

Variable system responses can occur between two system designs because of differences in:

System Lag – Time spent by robot processing information either after a command or sensory data is received. Example: Robot computing collision-free path when waypoint is requested.

Update Rate – Delays in displaying information to an operator due to variations in the size of the data or the capture rate of the hardware

Potential Sources of Biases in HRI – User

Factors that can influence human performance include:

1) **Operational factors** – Time deployed or geographic location

2) **Equipment factors** – Physical parameters (e.g. size, shape, color) or workspace layout

3) **Task factors** – complexity and repetitiveness

4) **Personnel factors** – Training, motivation, stress

5) **External environmental factors** – Temperature, noise level, visibility

6) **Role of the human** – Supervisor, operator, mechanic, peer, bystander

Task Metrics - Navigation

1) **Global Navigation** – Robot needs to know in general where it is located. Whether within a region, room, and floor in a building

2) **Local Navigation** – This focuses on the low-level navigation capabilities for smoothly navigating an environment. The robot should know what hazards are nearby such as stairs, doorways, pedestrians, or trees.

3) **Obstacle Encounter** – A robot should also be capable of navigating out of problems such as being stuck in debris or creating plans to address the potential hazards

Task Metrics - Navigation

Potential measures for navigation task performance/effectiveness:

- Percentage of navigation tasks completed
- Total area covered
- Deviation from planned routes
- Obstacles successfully avoided
- Obstacles not avoided but could be overcome

Potential measures for navigation task efficiency:

- Time to complete the task
- Operator time on task
- Average time to address an obstacle

Task Metrics - Navigation

Potential measures for navigation task human intervention (unplanned looping) and workload:

- # of operator interventions per unit time. An intervention refers to an unplanned interaction from the human.
- Average time needed for an intervention
- Ratio of operator time vs. robot time

Task Metrics - Perception

- Perception is the process of making inferences on distal stimuli (objects in the environment) based on proximal stimuli (energy detected by sensors).
- This can be achieved by both the **Human** and **Robot.**
  - Robot inference could include localization, obstacle detection, human recognition
  - Human inference could be identifying a victim from an image
  - Combination of both could be robot directing human attention to area of interest and human verifying a victim.
- There are two primary tasks for perception: 1) interpreting sensed data and 2) seeking new sensor data
- There are two categories for perception: 1) passive perception and 2) active perception

Task Metrics - Passive Perception

1) Passive perception – Interpreting sensor data for identification, judgement of extent, and judgement of motion tasks

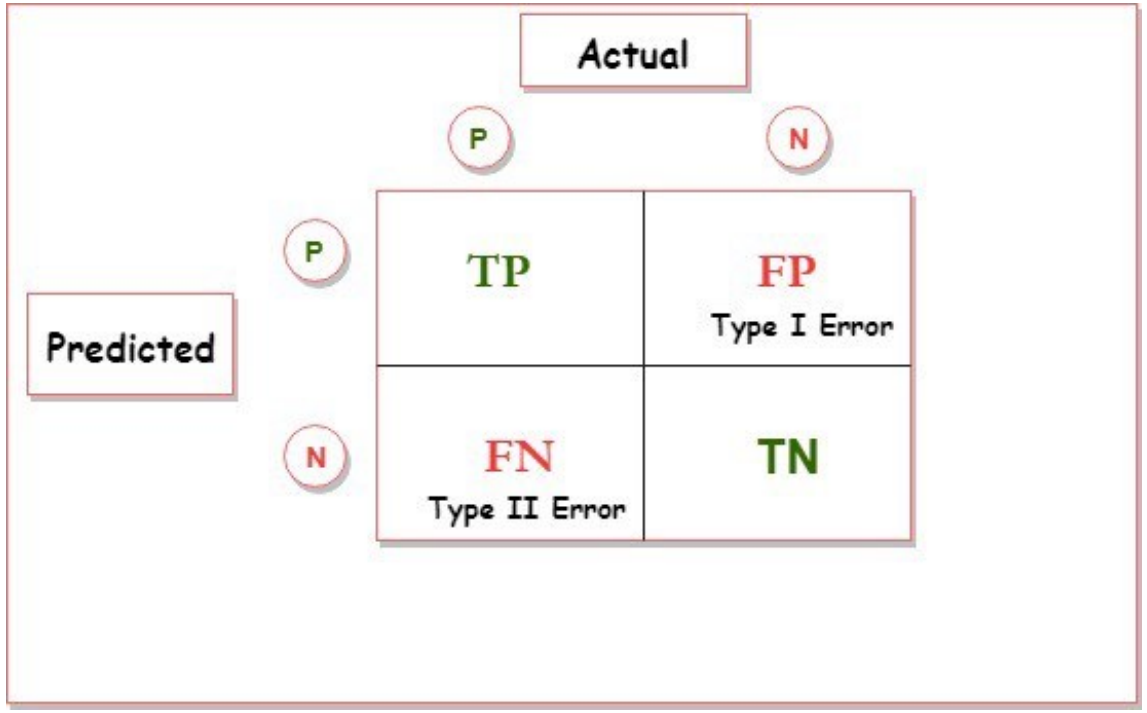
Identification – Detection and recognition of objects within sensor range. Detection vs Recognition?

Measures:

- Detection measures: % detected, signal detection, detection by object orientation, detection based on cluttered or sparse environments
- Recognition measures: Classification accuracy, confusion matrices, or recognition by object orientation

Precision = TP/(TP+FP) = The ratio of correct positive predictions to the total predicted positives

Recall = TP/(TP+FN) =Ratio of correct positive predictions to the total positives examples



Task Metrics - Passive Perception

Judgement of Extent – A human's ability to accurately make quantitative spatial judgements of the environment. This is often effected by viewing angle, heights, and field of view of robot's camera.

Measures:

- Absolute judgements of distance, size, or length
- Relative judgements of distance, size, or length
- Platform relative judgements of spatial information

Task Metrics - Passive Perception

Judgement of Motion – A human's ability to accurately judge their egomotion or movement of objects in the environment.

Measures:

- Absolutely velocity of robot
- Estimates of relative motion with moving object.

Task Metrics – Active Perception

2) Active Perception – Interpreting sensor data while sensor or robot is in motion.

Active Recognition – Recognition tasks where mobility of either a sensor or robot is required.

Measures:

Efficiency: Time to confirm recognition or improvement in recognition over initial detection

Effort: Amount of camera movement required to complete recognition

Task Metrics – Active Perception

Stationary search – Search tasks that do not require mobility and only pan/tilt of sensor or data fusion between sensors

Measures:

- Detection accuracy
- Time to search or non-overlapping coverage
- Coverage as percentage of potential sensor coverage
- Operator confidence in sensor coverage

Task Metrics – Active Perception

Active search – Search task requiring mobility. Task is initiated by potential objects within sensor range that may conceal a target.

Measures:

- Efficiency – time and effort used
- Identification errors: # of incorrect targets, number of targets misked
- Degree of operator fusion which refers to how well a system supports a users ability to assess remote scenarios (i.e. operators ability to utilize information from multiple sensors to attain situation awareness)

Task Metrics – Management

1) Fan out: Number of robots that can be effectively managed by or controlled by a human.
   - Useful for logistical demands of deployment, difficulty in managing a robot during use, and cost-benefit of a robot
   - Helps identify the upper workload of a human operator
2) Intervention Response Time: The time when an operator first recognizes a problem or robot first requires assistance
   - Delay occurs during supervisory control because operator does not devote full attention to the robot
   - Response time can be subdivided into (1) time to deliver request from robot, (2) time for operator to notice request, (3) situation awareness and planning time, and (4) execution time
3) Level of Autonomy Discrepancies: The ability of an operator to identify the correct level of autonomy to utilize
   - Different levels of autonomy suitable for different tasks, environments, and events.
   - Anecdotal evidence has shown that if an operator switched to autonomous in the correct time robot failures may not have happened
   - One way to measure this is to identify the optimal autonomy state for given scenarios and test users for choosing this state.

Task Metrics – Manipulation

1) Degree of mental computation
   - Measured by mental computation tasks such as mental rotations, rate tracking, and object-referent association.

2) Contact Errors
   - Measured as the # of unintentional collisions between a manipulator and the environment
   - Also can measure the type of contact errors (e.g. hard/soft touches)
   - For example, joint vs. end-effector control can lead to different contact errors especially in confined spaces.

Task Metrics – Social

1) Interaction Characteristics – Measured from observation of human behavior or conversations between agents

2) Persuasiveness – The ability of a robot to change the attitudes or behaviors a human
   - Measured by percentage of times a robot can alter a user's actions

3) Trust – Reliance on autonomation to perform a task. Especially in complex and dynamic environments where automation may not be perfect.

4) Engagement – A system's ability to maintain a user's attention
   - Measured by ability to capture attention and holding that attention.

5) Compliance – The amount of cooperation provided by a human to a robot.
   - Measured by percentage of times human responds to robot's requests

Common Metrics – System Performance

1) Quantitative Performance - Assessing the effectiveness and efficiency of team accomplishing a task

   Effectiveness – Percentage of task accomplished with **desired autonomy** level or number of operator intervention required

   Efficiency – Time required to accomplish a task. Can be broken down into all tasks completed or for only tasks completed by autonomy design.

2) Subjective Ratings – Assesses the quality of the effort according to stakeholders
   - The primary task may be finding a victim but attaining the quality information maybe important to the medical and structural engineering teams.

3) Appropriate Utilization of Mixed-Initiative – Ability of the human-robot team to regulate control initiative
   - Percentage of requests for assistance by robot
   - Percentage of requests for assistance by operator
   - # of interruptions of operator rated as non-critical

Common Metrics – Operator Performance

1) Situational Awareness

   Situational Awareness Global Assessment Technique (SAGAT) – Utilizes operator to queries to understand their situational awareness of the present context

2) Workload

   NASA – Task Load IndeX (NASA-TLX) – Widely used subjective measure for workload during robot operation.

   Physiological measures have also been investigated

3) Accuracy of mental models of device operation - conceptual, movement, spatial, and modality compatibility

   - The benefits of matching interface displays and controls to human “mental” models include reductions in mental transformations of information, faster learning and reduced cognitive load

Common Metrics – Robot Performance

1) Self-Awareness The degree to which a robot can accurately assess itself will have a direct impact on the ability of the human to efficiently interact with the robot. The less a robot is aware of its capabilities and the less it is able to recognize when it is having trouble, the more human monitoring and intervention is required.

   To qualitatively measure self-awareness, we propose assessing the following robot characteristics: (1) understanding of intrinsic limitations (mobility, sensor limitations, etc); (2) capacity for self- monitoring (health, state, task progress) and recognizing deviations from nominal; and (3) effectiveness at detecting, isolating, and recovering from faults (during both planning and execution).

2) Human-Awareness

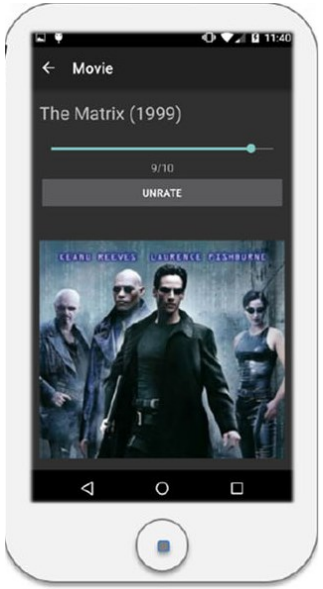
   Human awareness implies competency in various skills, the proficiency of which can be assessed independently or collectively. These include: (1) human-oriented perception (human detection and tracking, gesture and speech recognition, etc); (2) user modeling and monitoring (cognitive, attentional, activity); (3) user sensitivity (adapting behavior to user, measuring user feedback, recognizing human state).

Common Metrics – Robot Performance

3) Autonomy - The ability of robots to function independently

Neglect tolerance directly measures how a robot's effectiveness declines when the human is not attending to the robot. In particular, it measures the amount of the time the robot can be neglected before performance drops below an acceptable level of task performance

Netflix: Robot or Mobile Phone?



Condition APP: APP with text and graphics

Condition NAO: Robot just talking

Condition ENAO: Expressive robot which maps pitch, speed, gestures, and LED colors to suggestions

Hypotheses

H1: a humanoid robot interacting through natural modalities will be considered by the user more engaging and will be better liked with respect to a commonly used application on a mobile phone.

H2: As a consequence of H1, the animated robot obtains a persuasive effect on the user choices. Hence, the recommendations provided by the robot should be more likely to be accepted.

H3: The embodiment condition only (NAO) will not be preferred to the APP condition.

H4: The personality of the subjects will have an effect on the acceptance rate and on the evaluation of the interaction.

Experimental Setup

1) User provides 10 movie ratings and personal information (gender, age, instruction level, familiarity with Android apps, familiarity with movies, and personality questionnaire)

2) 6 recommendations created based on user profile and assigned to one of the 3 conditions

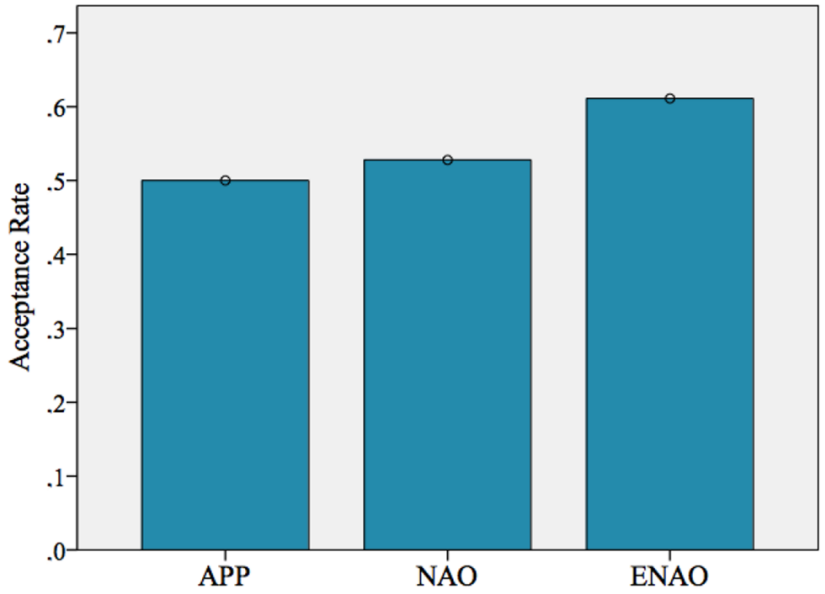
3) User completes a satisfaction and usability questionnaire

4) User asked to choose preferred interface

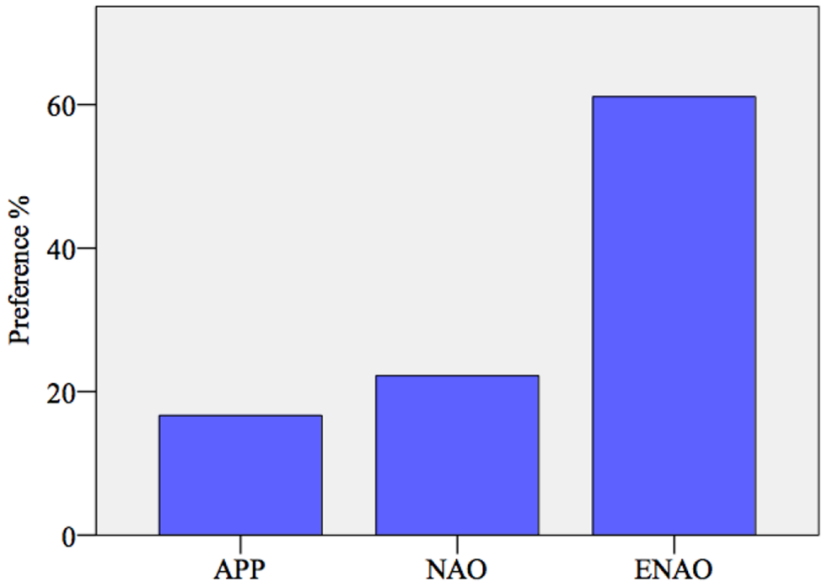
5) Recommendation acceptance was also tracked

Experiment 1 and Results

- Within-subject repeated measures experiment with English speaking robot
  - Independent var: Interaction condition (APP, NAO, ENAO)

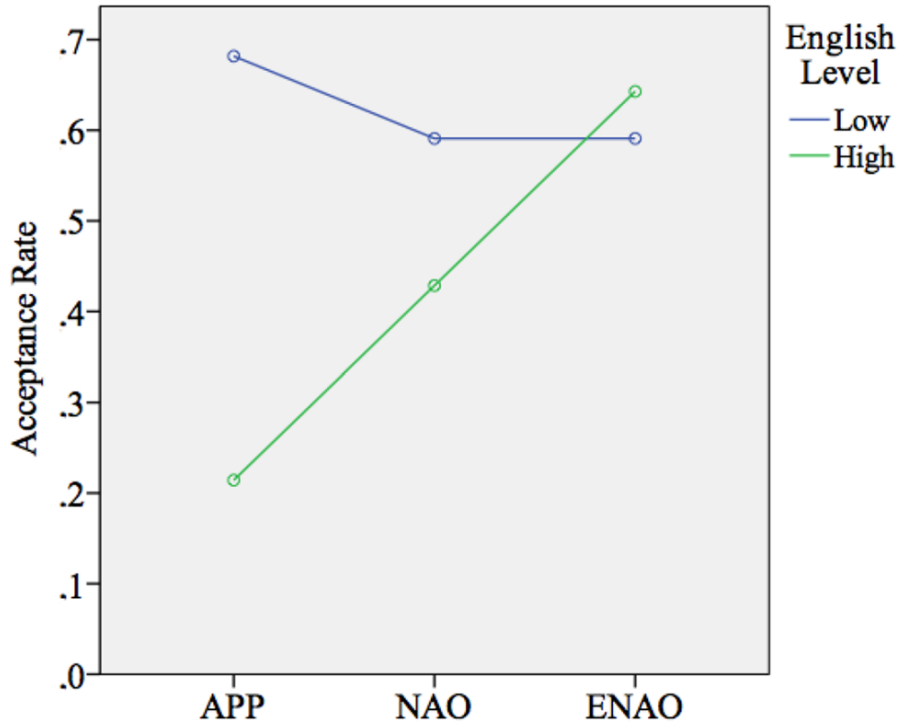


There is a minor difference between acceptance but not statistically significant



ENAO was preferred more often with statistically significant difference.

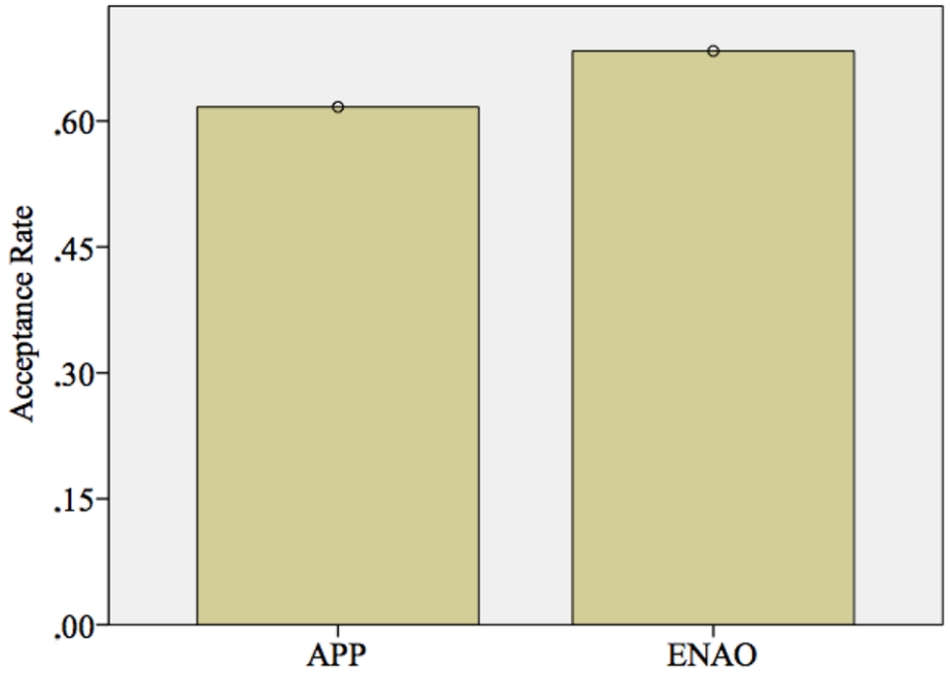
Experiment 1 and Results

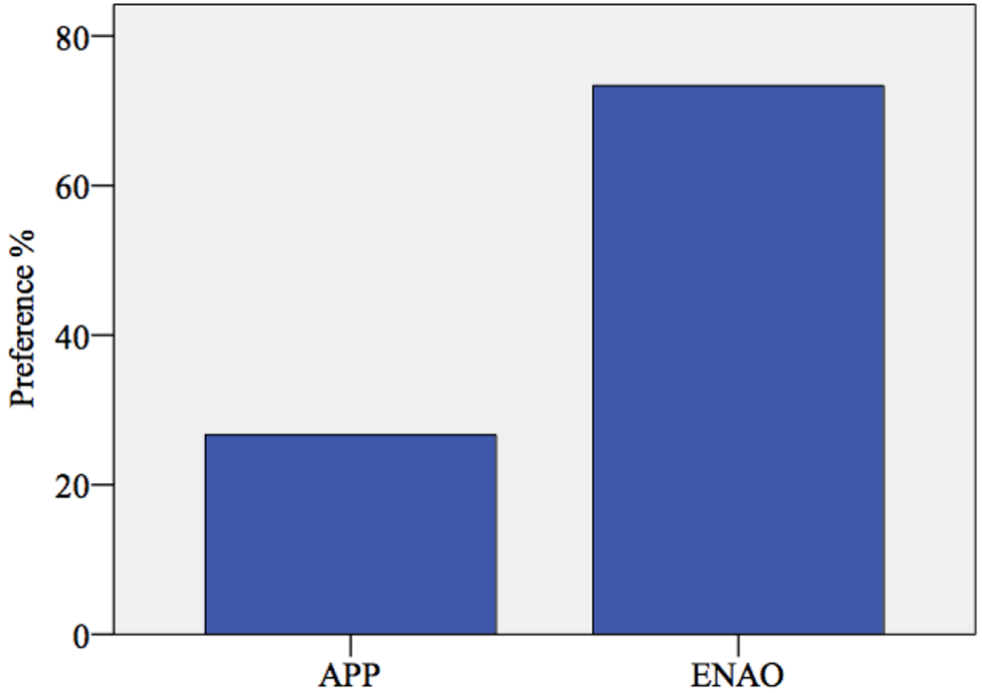


1) Higher movie acceptance rate with lower English proficiency
2) ENAO provides similar acceptance in both cases meaning it can address wide range of prociency levels

Experiment 2 and Results

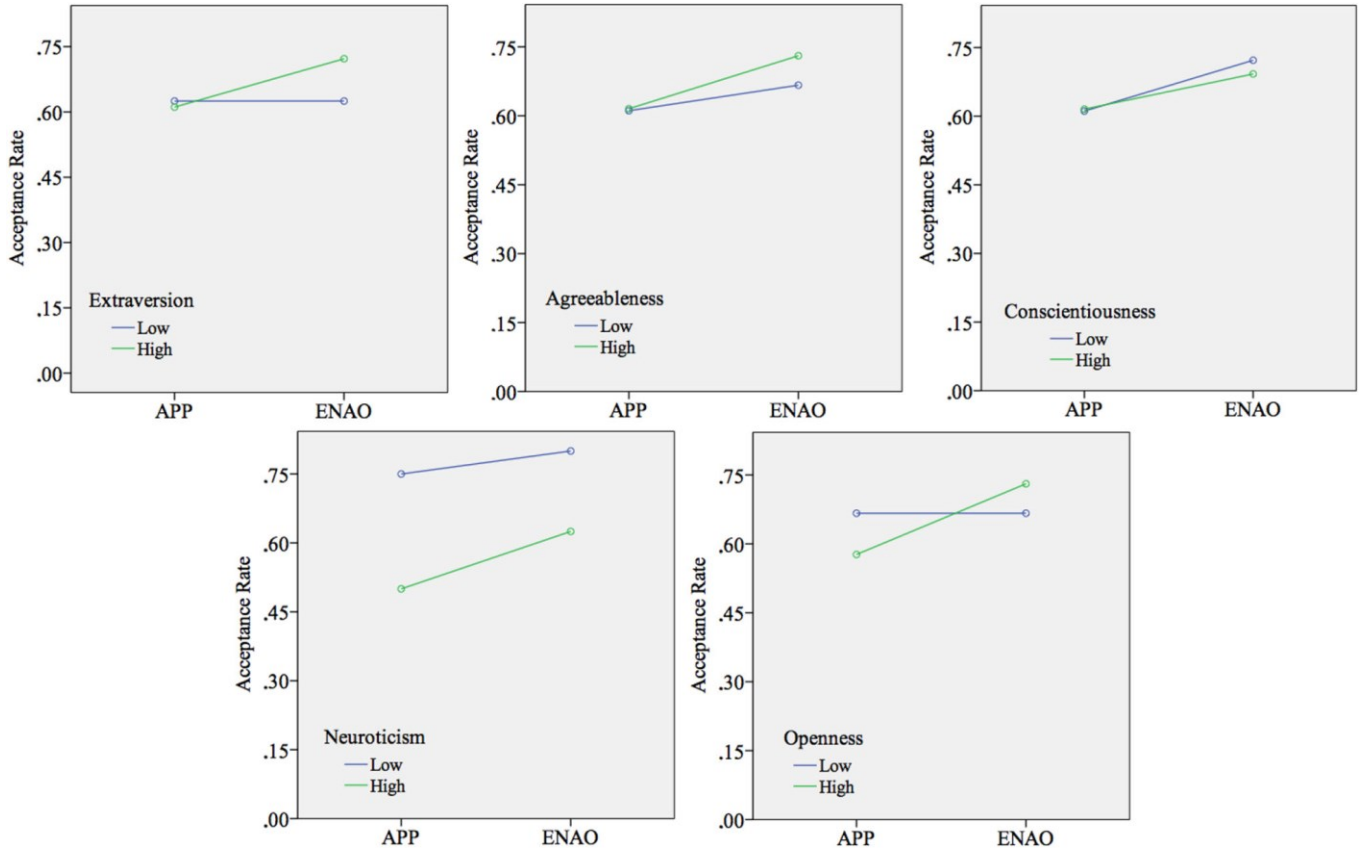
- Between-subject repeated measures experiment with Italian speaking robot





There is a minor difference between acceptance but not statistically significant

Experiment 2 and Results



Limitations?

Novelty Effect?

Response Bias?

Confounding Factors?