HRI Evaluation

Jean Scholtz, NIST

IEEE-RAS/ IFRR, 2004

Introduction
HCI compared to HRI

A Little HCI Background

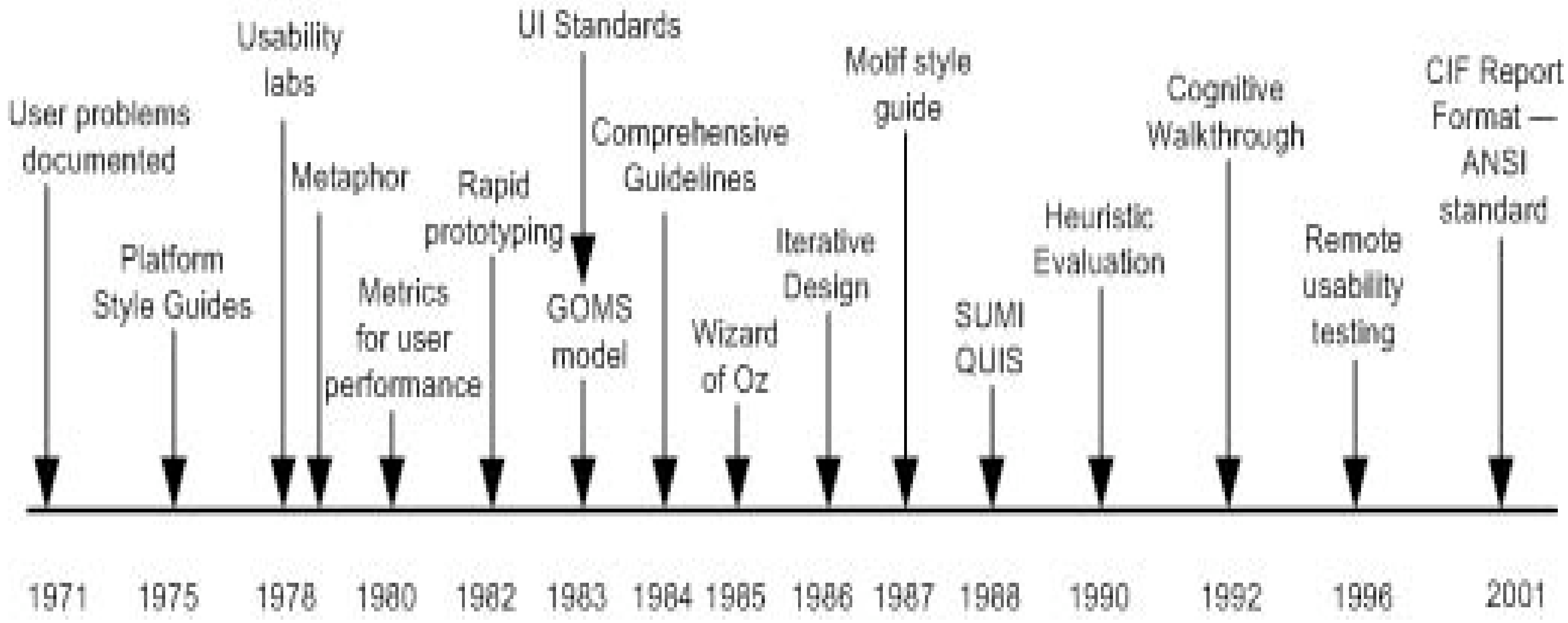
- Evaluation of human-computer interaction has focused on usability
- ISO 92407-11
  - Effectiveness: The percentage of tasks that users can complete, unaided, in a certain amount of time
  - Efficiency: The speed with which users can complete that task
  - User Satisfaction: How well users like using the software

  - The weighting given to the three metrics depend on the application
  - For example, discretionary/entertainment applications should score high in user satisfaction
  - Time critical applications should score well in efficiency
  - Life critical applications should score well in effectiveness

Different types of HCI Evaluations

- Model based
  - GOMS (Goals, Operators, Methods, Selection Rules)
  - Cognitive Walkthrough
- Empirical
  - Formative and Summative, usability testing
- Heuristic evaluations
  - Expert evaluations based on “rules of thumb”
  - Abstracted from guidelines
- Guideline based
  - “check-list” of certain guidelines

Evolution of HCI Evaluation Methods



30 years of highlights in the development of desktop computing user evaluations from 1971 - 2001

HCI vs. HRI

- HRI focuses on the combination of the user and the robot; HCI focuses on the user using the computer
- Different roles of interaction are possible
  - Roles discussed later
  - For now, different roles mean different informational needs and different interactions
- Multiple people can interact in different roles with same robot
- Users might interact with multiple (heterogeneous) robots at the same time Robot can act autonomously based on perception of world
- Degraded state of operation of robot
- Physical world – air, land, and sea!
- Intelligent systems, learning, emerging behaviors
- Harsh environments
  - Why do we send in robots?
- The user and the robot are not necessarily co-located
- Robots are evolving; tasks we want them for are evolving

Current State of HRI Evaluation

- We don’t have a good idea what good user interfaces/interactions for HRI really are
- We haven’t as yet identified good metrics/measures for HRI
  - Effectiveness, efficiency, and user satisfaction tell us only result
  - Need to understand why errors occur
- Methodology is currently limited to empirical testing as we don’t have guidelines and models necessary for developing other methods
  - But empirical testing is costly; even running subjects through simulations
  - Can never get the “same” conditions even in laboratory when using real robots; difficult to take the same path twice

HRI Evaluation Approach

- Determine what good metrics are for evaluation
- Based on this, work on developing methodologies for evaluation
- Start developing prototypes for user interfaces and interactions
  - And evaluating
  - And iterating
- Based on this work, develop guidelines and standards

- Questions
  - Effects of domains on guidelines
  - Different information /interactions based on roles?
  - How to accommodate different levels of autonomy?

Roles of Interaction

Roles of Interaction

- Why roles?
  - In HCI there is only the notion of the user
  - Not exactly true, there is also the programmer – who is rarely thought about in regard to user interfaces (7 volumes of Empirical Studies of Programming)
  - There is also the notion of system administrators; many programs have user interfaces for the “backend” as well.
- In HRI, needs of all users are not the same:
  - Supervisor
  - Operator
  - Teammate
  - Mechanic/Programmer
  - Bystander

Supervisor

- Oversees a number of robots
- May or may not have time to help one out
- May have to hand off to an operator
- Needs global picture of all robots/mission
- Needs to understand when a robot is having a problem, the seriousness of the problem, the effect on the mission

- Challenge:
  - How many robots can a supervisor effectively monitor?
  - How to show the robots in relationship to the global mission?
  - How to decide whether to hand-off a problem with a robot to an operator?

Operator

- Needs to have “telepresense” to understand where robot is and what must be done
- Interactions depend on level of autonomy
- Can vary from complete teleoperation to giving new way points to giving high level task to specifying a mission
- Needs awareness of robot health, awareness of environment and awareness of what robot is to be doing to support task/ mission

- Challenges:
  - How to maintain awareness despite communications limitations?
  - How to control multiple robots?
  - How to acquire situation awareness if attention has been diverted for some time?

Teammate

- Robot is a member of the team
- Teammates can give commands within the scope of the task/ mission
- Interactions such as gestures and voice may be helpful here
- Need to understand any limitations robot has in capabilities

- Challenge:
  - Can the robot understand the same interaction vocabulary as other team members?
  - How to understand any degraded conditions?
  - Can multiple teammates interact? How to know what previous tasking was?

Mechanic/Programmer

- Comes into play if the operator cannot resolve the issue
- These interactions could happen within a task or mission
- Given that a hardware/ software change is made, then the mechanic/programmer must have a way of interacting with the robot to determine if the problem has been solved.

- Challenges:
  - How much self diagnosis can the robot do?
  - How to determine when to move from operating in degraded capability to pulling robot off task and attempting to fix problem ?

Bystander

- No formal training using robot but must co-exist in environment with robot
  - Consider health care situation; floor cleaning robots; robot pets; on-road driving
- In military situations, could be a friendly, a neutral or an enemy
  - The robot should be able to protect itself from an enemy

- Challenges:
  - How can a bystander form a mental model of what the robot's capabilities are?
  - Should a bystander have a subset of interactions available?
  - What type of social interactions come into play?

Caveats to Roles

- One person might be able to assume a number of roles for a particular robot (excluding the bystander role)
- A number of people might be interacting with one robot in different roles; these people may have to be aware of the different interactions happening as well as other information they need.
- Not all systems will have / need all the roles
- Assuming we can determine information/ interaction needs for different roles, then we could use that information to:
  - Design a user interface to support a given role
  - Determine whether multiple roles could be supported in one user interface

Where to Study HRI?

Field Studies

- Advantages:
  - Good to understand the current constraints and the workflow
  - Identification of good metrics is possible in field studies
    - What is really important to people
  - Technology has to be reasonably robust to study in this environment
  - You will find out issues that have to be more deeply investigated in other environments
- Disadvantages:
  - Little control possible
  - What you see is current practices: inserting new technology will change this practice

Laboratory Studies

- Advantages:
  - Can control studies and isolate variables to determine effects
  - Can conduct studies with early prototypes of systems
- Disadvantages:
  - Difficult to simulate real-world conditions
  - Need to make sure that what you learn is valid in real-world
  - Using real robots make it difficult to replicate conditions

Simulations

- Advantages:
  - Less costly than field studies
  - Don’t need physical hardware (a real advantage!)
  - Can simulate wide variety of conditions that may not be seen in field test, or duplicated in laboratory
- Disadvantages:
  - Outside constraints (pressure, environmental conditions) can’t be duplicated

Example

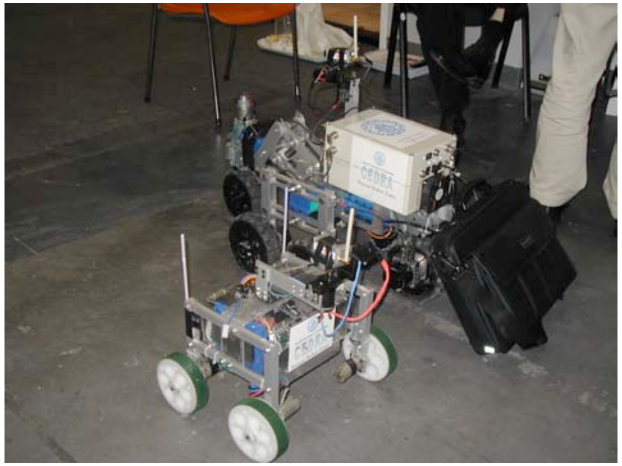
- We have used the NIST Test Arena to study USAR HRI
- http://www.isd.mel.nist.gov/projects/USAR/

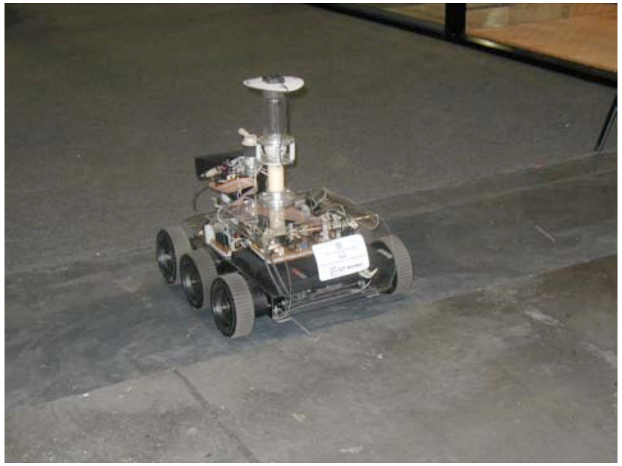
- Advantages:
  - Wide variety of user interfaces; levels of autonomy
  - Same conditions for all the teams
  - Can see progress from year to year
- Disadvantages:
  - We have no control over UIs, users, what users choose to do

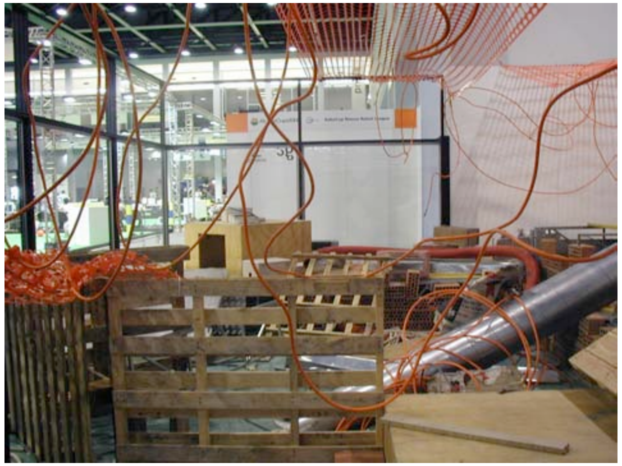
Urban Search and Rescue Test Arena

- Locate as many victims as possible with as few penalties as possible
- Score is determined by # victims (difficulty of arena located in), penalties (bumping into walls, victims), number of operators per robot, accuracy of maps of victim location

Search and Rescue Competitions







Photos are by permission of RoboCup 2003 teams

Video from Robocup 2003

Metrics

Proposed Metrics

(discussed in methodology section)

- Domain Specific
  - USAR competitions
    - Time spent navigating, overhead, obstacle extraction, logistics
    - Amount of space covered
    - Critical incidents
      - Positive outcomes
      - Negative outcomes
  - On-road driving
    - Simulations – SAGAT
    - Workload
  - Off-road driving
    - Field study
    - Operator interventions – time, time to acquire SA, reason for intervention
    - Workload

Proposed Metrics

- Domain Independent (Goodrich, Crandall, Oslen)
  - Neglect tolerance: a measure of the effectiveness of the robot's autonomy mode. How is the robot's performance affected when neglected for a period of time?
  - Interface efficiency: time it takes the human to gain Situational awareness; time to formulate a plan, time to translate that into commands, time to communicate that plan to the robot
  - Tolerance of world complexity: how well does the interaction scheme scale to degrees of complexity in the world?
  - Robot Attention Demand: average time spent servicing the robot/ divided by sum of average time spent servicing robot + neglect tolerance
- Workload
- Efficiency (GOMS analysis)?
- Accuracy of mental models

Proposed Metrics, cont.

Performance of an interaction scheme

$p = V(\pi; t, c, t_N)$ where π is the interaction scheme

$c = C(s)$ where C is a complexity metric and s is a set of states

C needs to consider branching factors for navigation and amount of clutter

Really two equations as time can be t_{on} or t_{off} depending on whether the robot is being serviced or not

$p = V_s(\pi; t_{on}, c, t_N)$ or $p = V_N(\pi; t_{off}, c, t_N)$

$p = V(\pi)$ is the average frequency and duration of interactions between robot and human

Using this Metric

Need to have humans and robot interact in real tasks. Use secondary tasks to thoroughly sample the domain space.

Need a large number of subjects to sample discrete events in the continuum

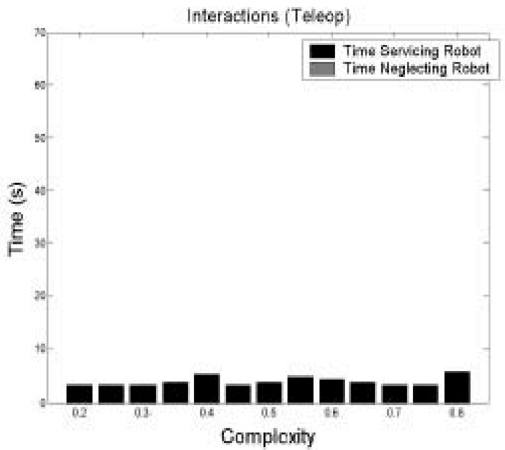
Complexity estimate is an issue – unknown domain

Use a Gaussian filter on the data according to the central limit theorem to lower the number of subjects needed

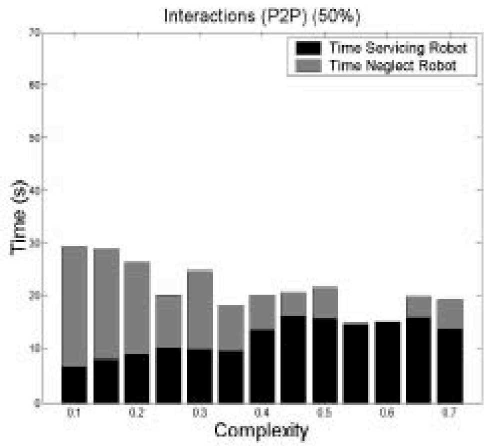
Example from Crandall and Goodrich

- Compared 3 autonomy schemes
  - Teleop, P2P, scripted
  - P2P: gave robot directions for next intersection
  - Scripted: set way points for robot to follow
  - 2 secondary tasks: navigate another robot; 2 digit addition and subtraction
  - Operator worked with robot, pushed a button when done. Did arithmetic tasks until time to service other robot
  - t_N has to be extended until for some times the effectiveness drops to 0. Thus t_N is different for all 3 conditions
    - 1 for teleop
    - P2P: t_N was divided into 5,10, 15, 20, 25, and 30 second bins
    - Scripted: t_N was divided into 10, 20, 30, 40, 50, and 60 second bins

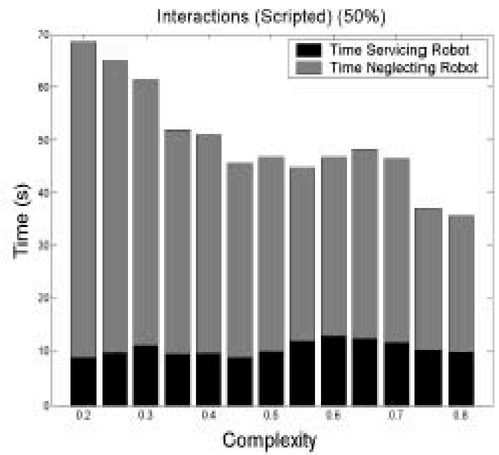
Results



Interactions (Teleop)
Time Servicing Robot
Time Neglecting Robot
Time (s)
Complexity

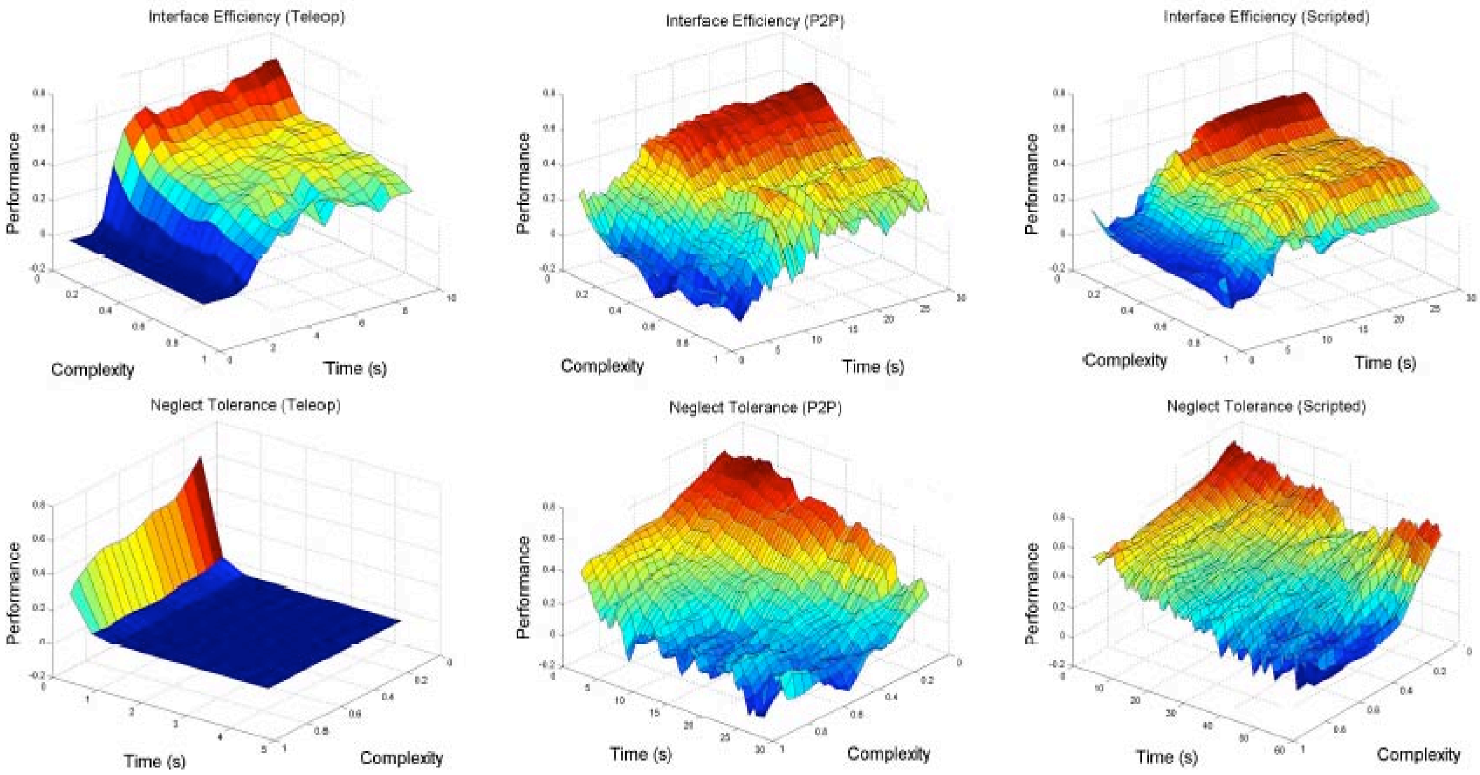


Interactions (P2P) (50%)
Time Servicing Robot
Time Neglect Robot
Time (s)
Complexity



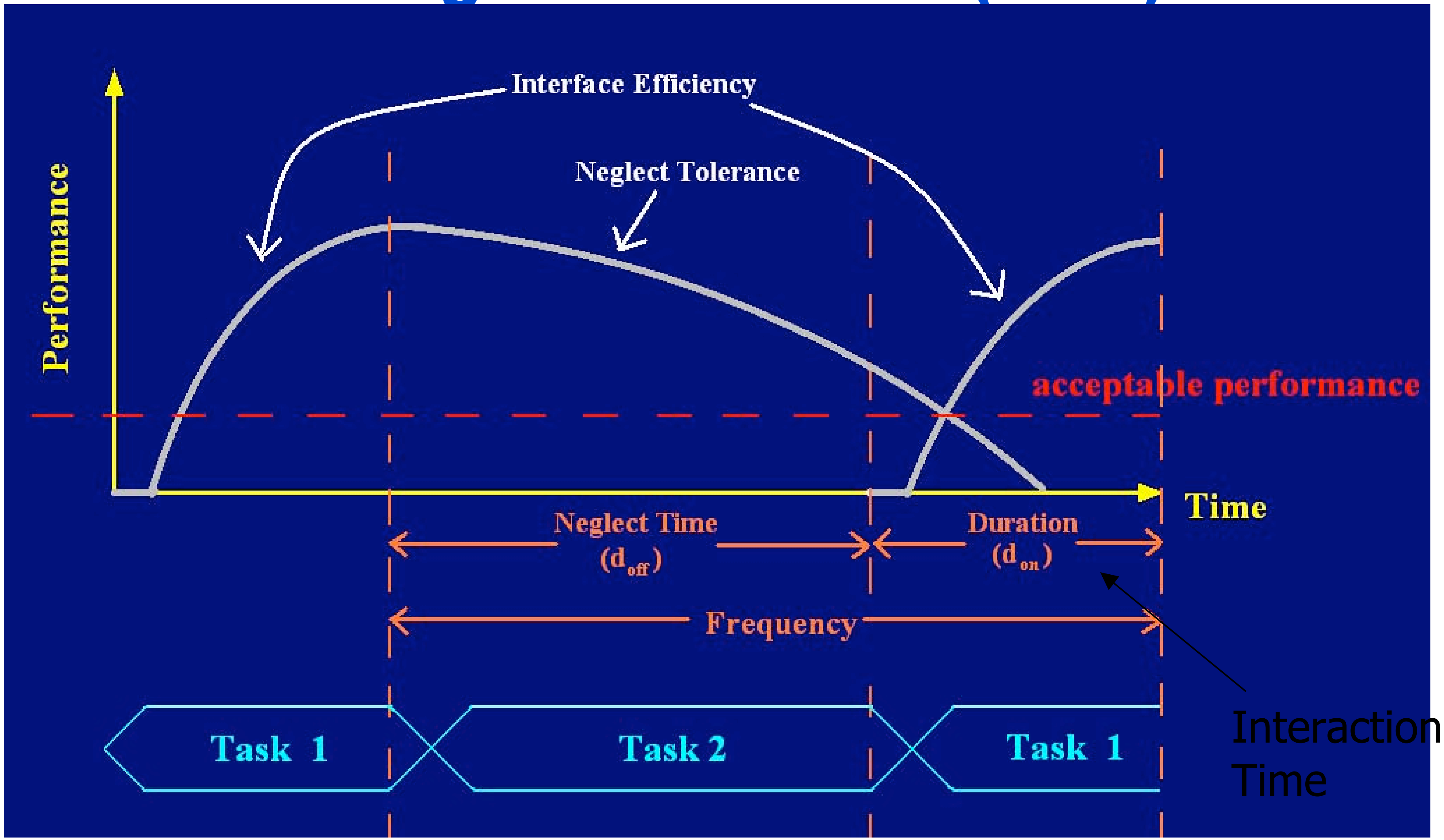
Interactions (Scripted) (50%)
Time Servicing Robot
Time Neglecting Robot
Time (s)
Complexity

Results



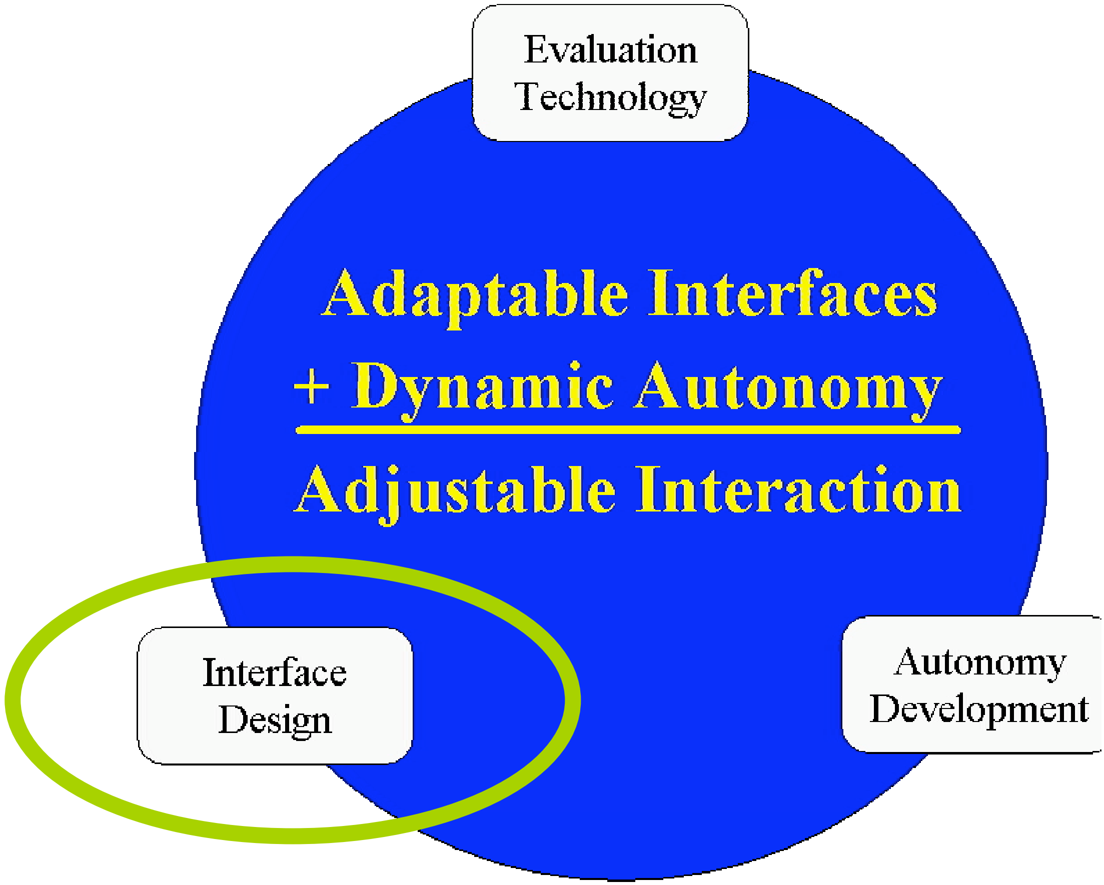
Plots of the mean of the random processes $V(Teleop)$, $V(P2P; t_N = 30sec.)$, and $V(Scripted; t_N = 60sec.)$.

Combining Interface Efficiency and Neglect Tolerance(BYU)



Interfaces: Principles of Efficient Interaction (BYU)

1. Implicitly switch
2. Use natural cues
3. Manipulate world
4. Manipulate relationship
5. Manipulate info
6. Externalize memory
7. Manage attention
8. Learn



Efficiency – GOMS Analysis

- From the Psychology of Human-Computer Interaction, Card, Moran, and Newell
- Based on a model of the human processing unit
- Perceptual system, motor system, cognitive system
  - Times can be identified for typical actions
  - Typing, retrieving an item from long term memory, eye movements, decay time for visual memory
  - Using these times and a model of how people recognize- act, we can compute average times for a human-system interaction at the keystroke level

Efficiency – GOMS Analysis

- User's cognitive structure contains:
  - Set of goals
  - Set of operators
  - Set of methods for achieving the goals
  - Set of selection rules for choosing among the methods
- Example in USAR
  - Goal: locate all victims
    - Goal: locate victim – repeat until no more victims
      - move robot through area
      - Identify any victim

GOMS, cont.

- Move robot through arena
  - Select: teleop, full autonomy, specify waypoint

- Goal: specify waypoint
  - Mark specified waypoint on map
    - Position cursor (locate cursor, locate designated point, move cursor to that point)
    - Double click

  - Select “enter waypoint” command
    - Determine coordinates of waypoint
    - Type in x coordinate
    - Type in y coordinate
    - Select enter

GOMS, cont.

- Each of these actions needs to be broken down into recognize-act cycle
- And times assigned

This allows you to determine efficiency of different types of interactions

BUT it does not provide any information about possible errors

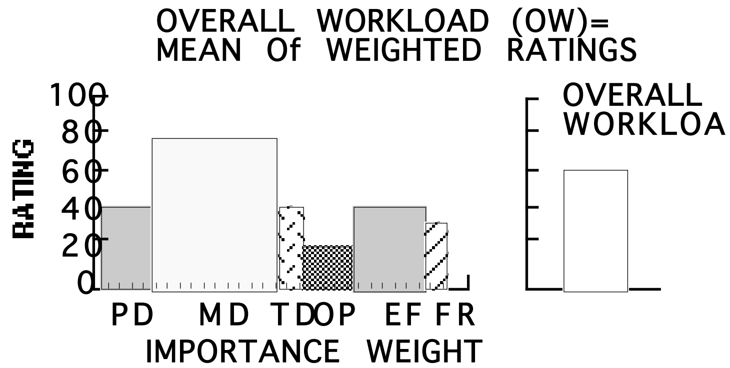
Workload Metric (Domain Dependent)

- Subjective ratings
- NASA TLX (Task Load Indicator) is the best known one for use in cognitive load
- 6 subscales
  - Mental Demands
  - Physical Demands
  - Temporal Demands
  - Own Performance
  - Effort
  - Frustration
- 15 possible pairings of the six subscales to determine which is the greater factor in the task
- Give these pairings to determine weightings

Assessing Workload, cont.

- Step 1 – determine weights
- Step 2 – determine ratings for each factor
- Step 3 – combine based on ratings/weights

- You need to determine a different set of weights for each distinctly different task
- Ratings are on a 20 point scale from low to high



RATING SCALE DEFINITIONS		
Title	Endpoints	Descriptions
MENTAL DEMAND	*Low/High*	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
PHYSICAL DEMAND	*Low/High*	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
TEMPORAL DEMAND	*Low/High*	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
EFFORT	*Low/High*	How hard did you have to work (mentally and physically) to accomplish your level of performance?
PERFORMANCE	*Good/Poor*	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
FRUSTRATION LEVEL	*Low/High*	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

Study- TRL-6

- Schipani, Permis 2003: Study of workload on operators (TRL 6 study) in 3 types of terrain (arid, wooded, urban)
- Sampled workload 5-10 second values during intervention segment
- Found that operator interventions significantly affected workload

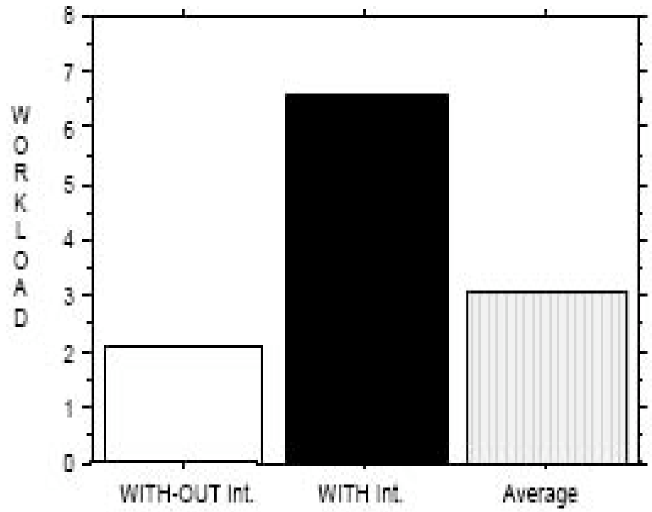


Figure 3. Plot of With-Out and With Intervention, including Average Workload, all test Phases.

Results

- Arid terrain
  - Mean workload without intervention 1.6 (10)
  - Mean workload with intervention 5.8
- Wooded terrain
  - Mean workload without intervention 2.4
  - Mean workload with intervention 7.75
- Urban terrain
  - Mean workload without intervention 1.9
  - Mean workload with intervention 6.1

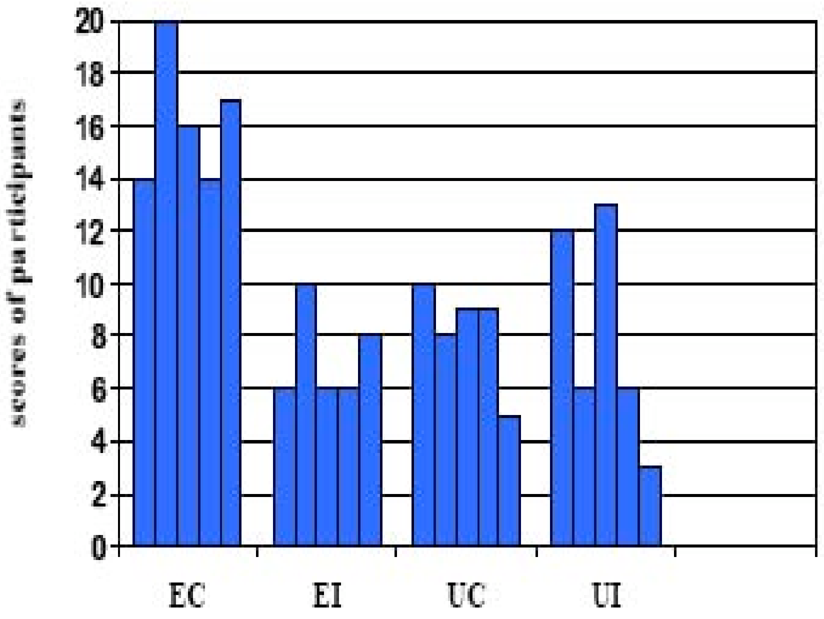
- Temporal, mental categories were highest contributors

Mental Models

- Can the user build up a mental model of the robot's interactions "bottom-up"?
- Study
  - Used an Aibo
  - Programmed both dog-like and un-doglike behaviors, consistent and inconsistent behaviors
  - Asked subjects to "play" with Aibo for 10 minutes
  - Asked them about various interactions and satisfaction

Mental Models: Results

- Metrics: predictability of behavior; capability awareness; interaction awareness; user satisfaction
- Capability awareness – asked prior to interaction
- Interaction awareness – see graph
- User Satisfaction – more doglike behaviors, multiple commands, difficult to tell when interaction was finished



Methodologies

Situational Awareness

- Endsely def. SA is the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future.
- Three Levels:
  – Level 1: perception of element
  – Level 2: comprehension
  – Level 3: projection
- Assessment method
  – Freeze simulation at certain point in time
  – Ask user questions about relevant information appropriate to task

Approach

User Side

- Determine information needed for various human roles; situations
- Incorporate into “system” architecture
- Develop a situational awareness assessment tool based on task analysis
- Prototype a user interface
  - assure usability of interface
  - develop SA metrics using SAGAT methodology
- Deliver HRI SAGAT / baseline architecture/ UI to community

Robot Side

- Determine information needed for_robot interaction
- Incorporate into “system” architecture
- Run simulations - before and after
  - Measure user workload; user performance; mission “success”
- Deliver augmented architecture to community

Situation Awareness: On-Road Driving

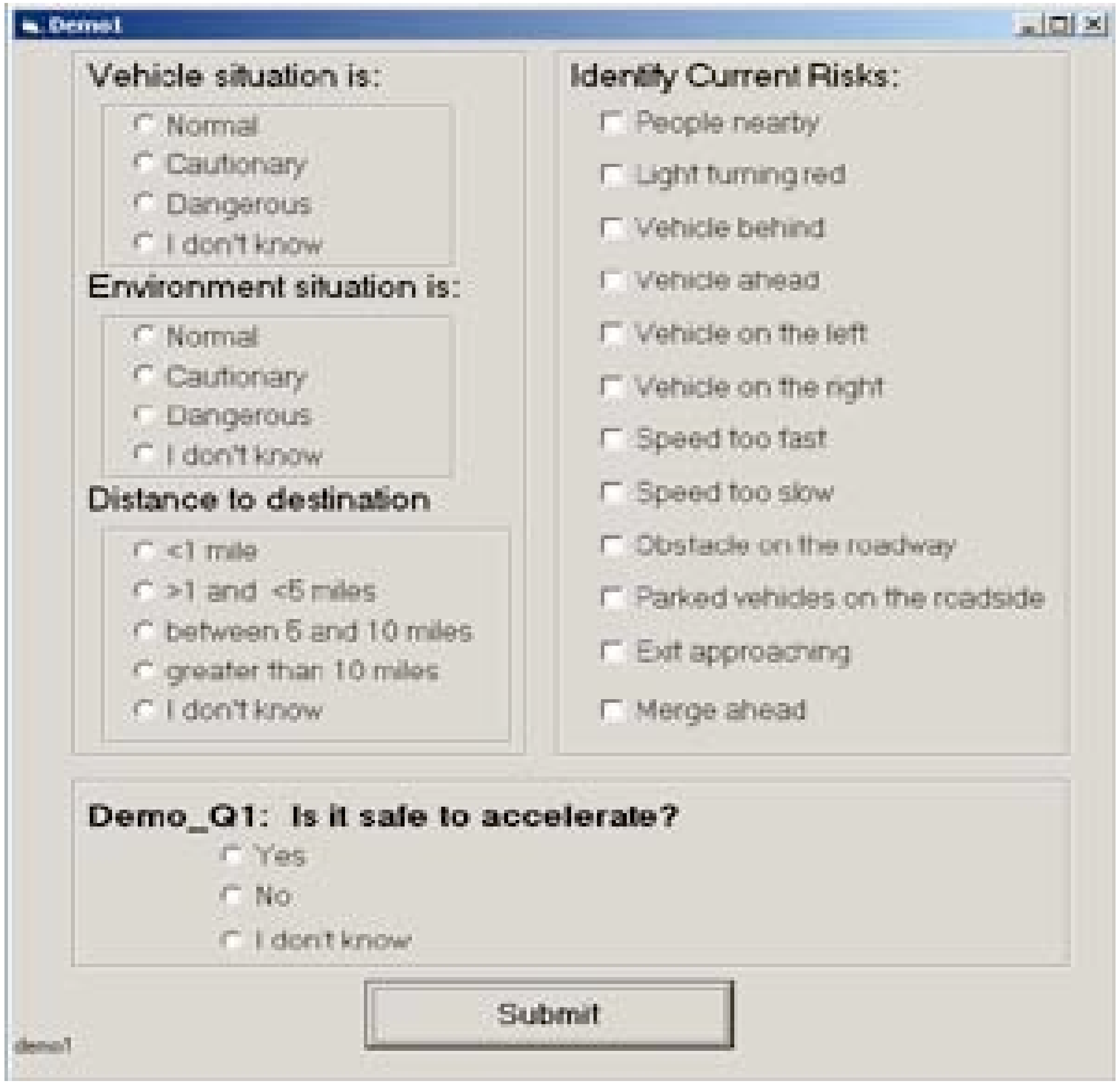
- Developed prototype GUI
- Developed questions for 3 levels of SA
  - Used driving program as “expert”
- Developed data collection interface for SA
- Developed simulation data for different conditions
- Conducted Experiment

- Iterated on experiment to clean up some ambiguous questions
- Now working on better GUI – to test sensitivity to GUI

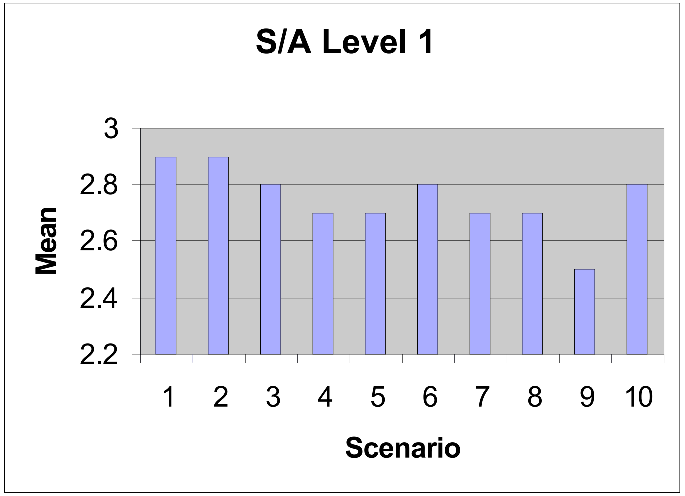

SimDrive v1.01 -- user01 - demo1
1x
George Washington Memorial Parkway
482yrds
Vehicle: Bravo
Speed: 50.2 MPH
Destination: Urbana, MD
Distance to Dest.: 1.69 mi
Current Road: GW Parkway
Latitude: 38.828300 N
Longitude: -77.043030 W
Vechile Detailed Inf
Vehicle Detail: Bravo
Speed: 50.2 MPH
Fuel: 1/4
Road: GW Parkway
Location: 38.828300 N
-77.043030 W
Lane: Right (2 lanes)
Behavior: Slowing - Congestion
Alerts: None
Close
Environment Detailed Info.
Vehicle: Bravo
Four Lane Divided
Posted Speed Limit: 45 MPH
Environment Detail: Bravo
Signs
None
Vehicles
Car (right-behind, speed=35, distance=10m)
Pedestrians
None
Obstacles
Debri (right lane, distance=15m)
Alerts
Road Blocked
Close

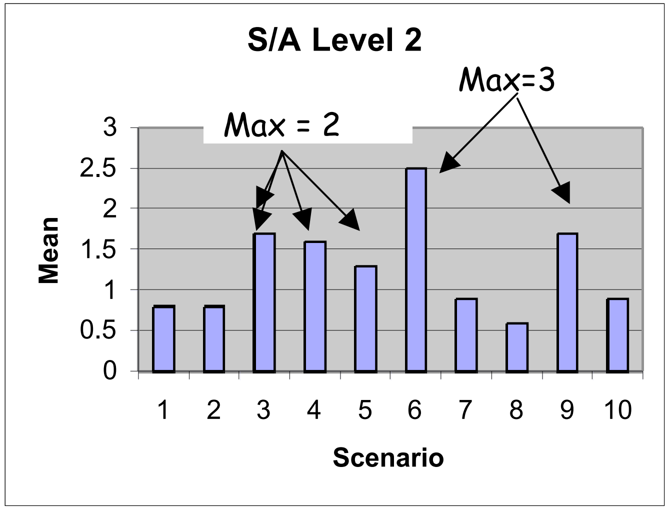
Situation Awareness Assessment

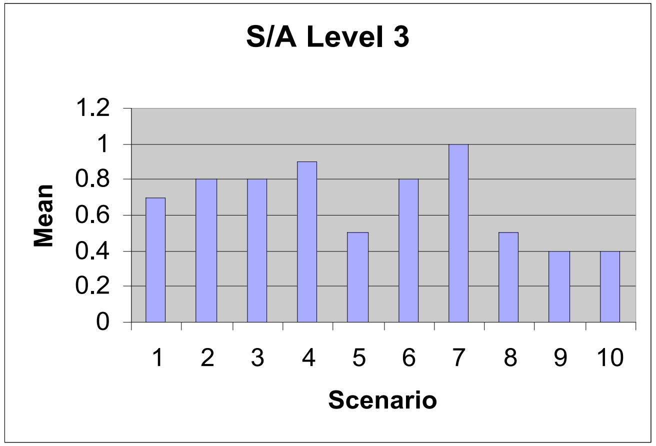
- Level 1 – perception
  - How far is it to destination? Is vehicle health okay?
- Level 2 – comprehension
  - What are potential risks?
- Level 3 – Projection
  - Can you turn left now? Is it safe to pass?


Demo1
Vehicle situation is:
Normal
Cautionary
Dangerous
I don't know
Environment situation is:
Normal
Cautionary
Dangerous
I don't know
Distance to destination
<1 mile
>1 and <5 miles
between 5 and 10 miles
greater than 10 miles
I don't know
Identify Current Risks:
People nearby
Light turning red
Vehicle behind
Vehicle ahead
Vehicle on the left
Vehicle on the right
Speed too fast
Speed too slow
Obstacle on the roadway
Parked vehicles on the roadside
Exit approaching
Merge ahead
Demo_Q1: Is it safe to accelerate?
Yes
No
I don't know
Submit
demo1

Results - SA


S/A Level 1
Mean
3
2.8
2.6
2.4
2.2
1 2 3 4 5 6 7 8 9 10
Scenario


S/A Level 2
Max=3
Max = 2
Mean
3
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10
Scenario


S/A Level 3
Mean
1.2
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10
Scenario

Results - Workload

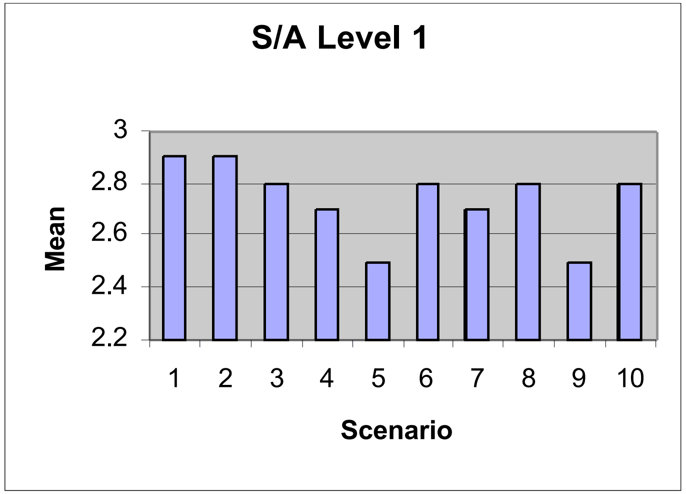
	Range	Min	Max	Sum	Mean	Std. Dev.
S 1	67	10	77	348	34.80	23.99
S 2	45	21	66	402	40.20	15.67
S3	69	8	77	341	34.10	21.01

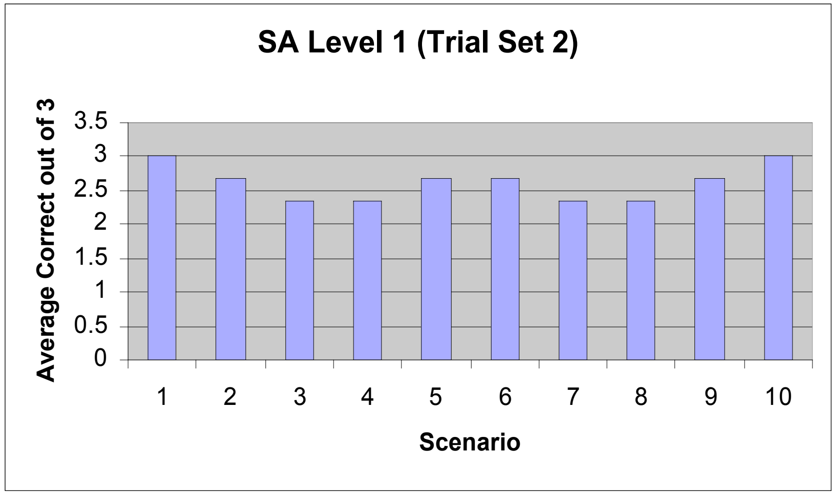
Second Iteration

- Cleaning up some ambiguous questions
- Reran experiment

SA level 1

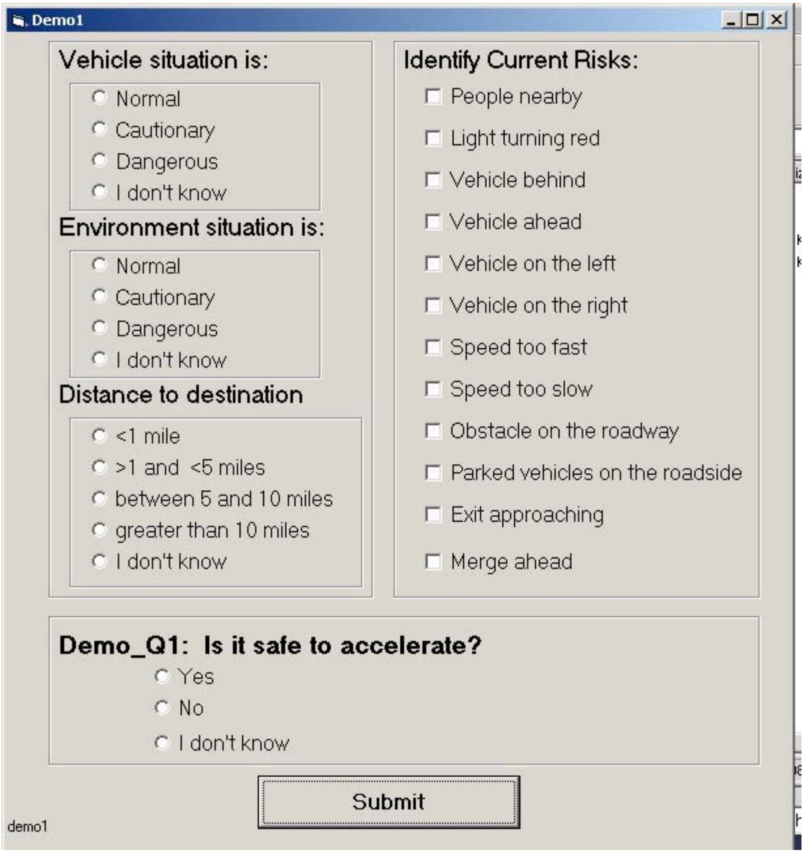
No changes made – approximately the same as before





SA Level 2 – Changed

Made the risks more precise but that added details



Demo1

Vehicle situation is:
- Normal
- Cautionary
- Dangerous
- I don't know

Environment situation is:
- Normal
- Cautionary
- Dangerous
- I don't know

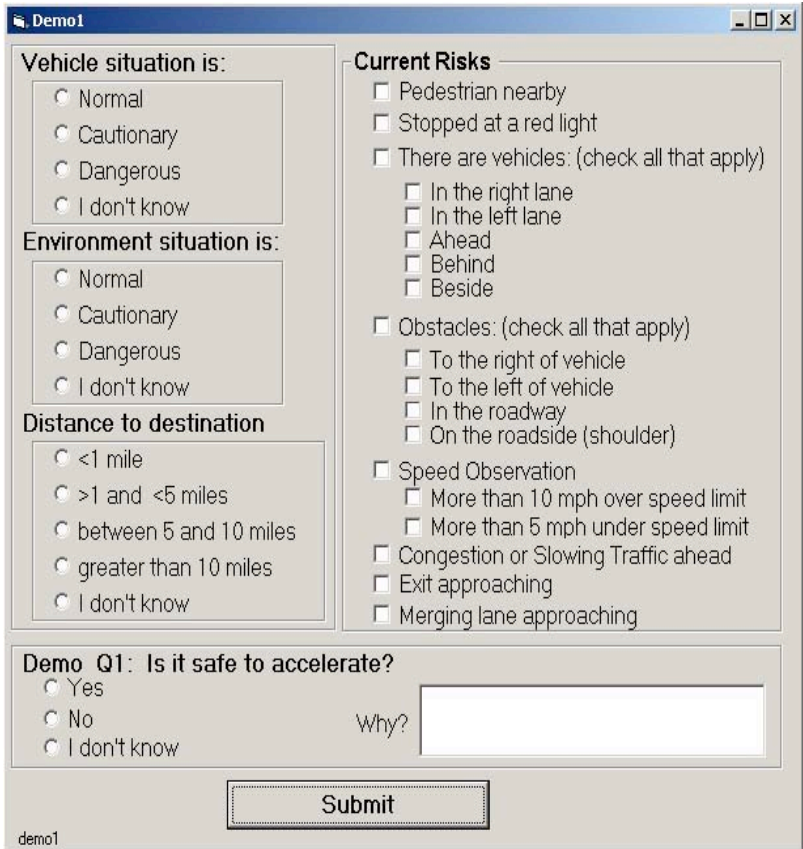
Distance to destination
- <1 mile
- >1 and <5 miles
- between 5 and 10 miles
- greater than 10 miles
- I don't know

Identify Current Risks:
- People nearby
- Light turning red
- Vehicle behind
- Vehicle ahead
- Vehicle on the left
- Vehicle on the right
- Speed too fast
- Speed too slow
- Obstacle on the roadway
- Parked vehicles on the roadside
- Exit approaching
- Merge ahead

Demo_Q1: Is it safe to accelerate?
- Yes
- No
- I don't know

Submit

demo1



Demo1

Vehicle situation is:
- Normal
- Cautionary
- Dangerous
- I don't know

Environment situation is:
- Normal
- Cautionary
- Dangerous
- I don't know

Distance to destination
- <1 mile
- >1 and <5 miles
- between 5 and 10 miles
- greater than 10 miles
- I don't know

Current Risks
- Pedestrian nearby
- Stopped at a red light
- There are vehicles: (check all that apply)
  - In the right lane
  - In the left lane
  - Ahead
  - Behind
  - Beside
- Obstacles: (check all that apply)
  - To the right of vehicle
  - To the left of vehicle
  - In the roadway
  - On the roadside (shoulder)
- Speed Observation
  - More than 10 mph over speed limit
  - More than 5 mph under speed limit
- Congestion or Slowing Traffic ahead
- Exit approaching
- Merging lane approaching

Demo Q1: Is it safe to accelerate?
- Yes
- No
- I don't know

Why?

Submit

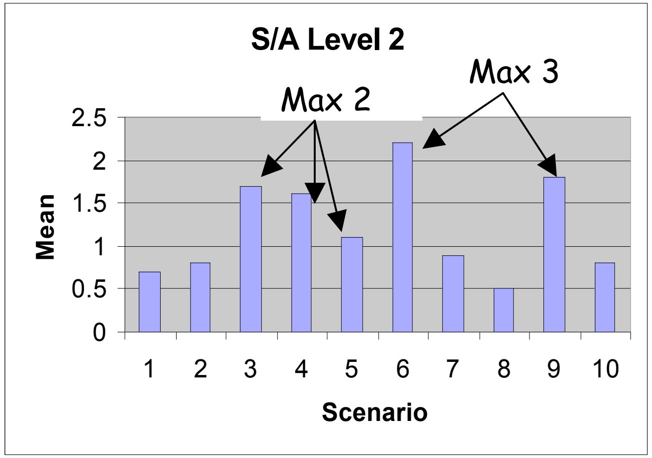
demo1

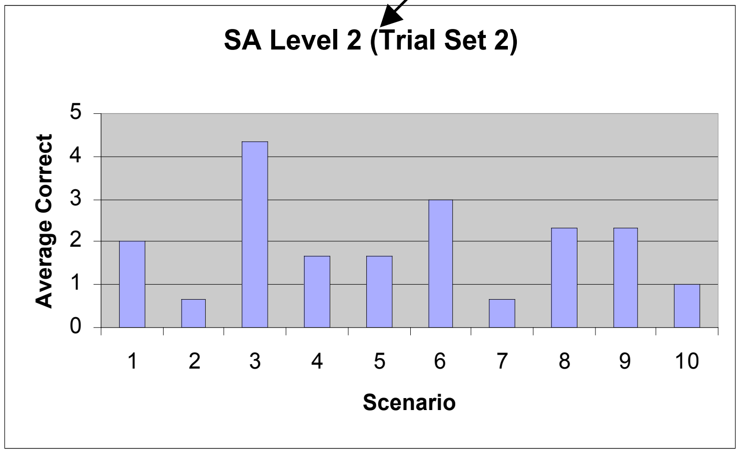
SA Level 2

For SA 2 (Max number correct per scenario):

1: 2	6: 4
2: 1	7: 2
3: 6	8: 7
4: 6	9: 7
5: 3	10: 1

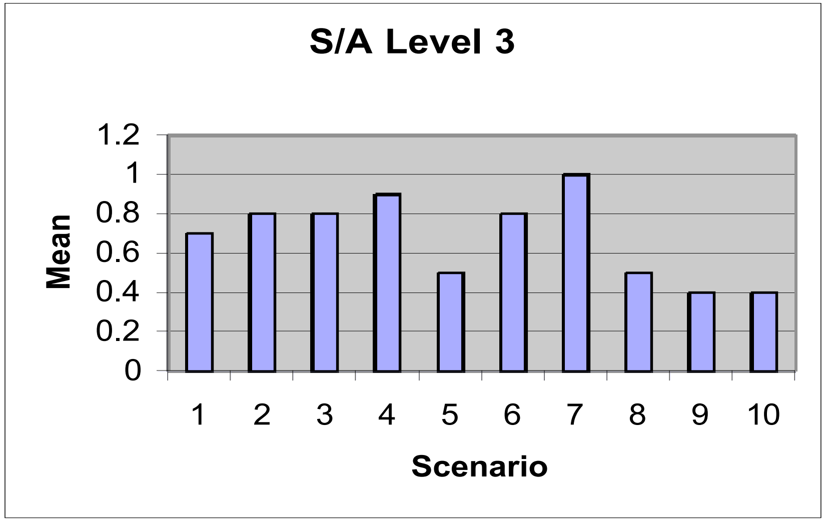
More details-> lower scores

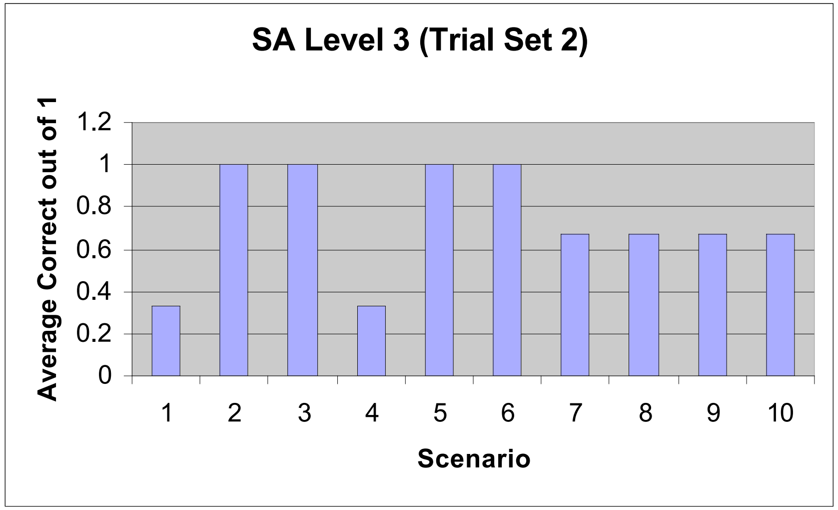




SA Level 3- Changed

Mixed results – also asked for rationale





Rationale

- Scenario 1: can vehicle legally turn left?
- Correct answer: no – vehicle is in right lane
  - No – stopped at red light
  - Yes – nothing to indicate it can't
  - Yes- no explanation
- Scenario 4: Is it save to swerve into left lane?
- Correct answer: No there is a car beside vehicle
  - 2 don't know
  - 1 no – car beside
- Currently in the process of recoding the UI and will take into consideration what information has to be comprehended to correctly answer these questions

Interventions: Off-road Driving

- Metrics used
  - Number of interventions
  - Type of interventions
  - Time needed for intervention
  - Percentage of intervention time needed for SA

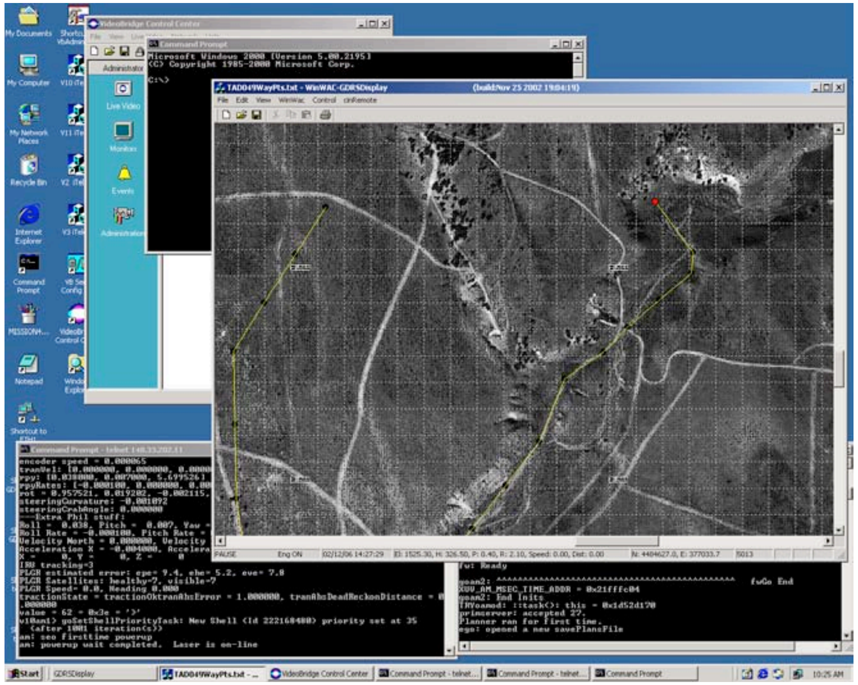
- Study
  - 3 types of terrain (two analyzed here)
  - Each study took about 10 days
  - 2 operators, 2 identical robots
  - Experiment looked at 3 different lengths of runs; easy and more difficult course (terrain features)

Trials

Number of trials	500 Meters	1000 Meters	2000 Meters	7000 Meters	Totals
ARID Gold/ LOS	12	21	12	1	91
ARID Gold/ NLOS	12	21	12		
WOODED Gold/LOS	12	21	12	1	91
WOODED Gold/NLOS	12	21	12		
ARID Black/ LOS	12	18	12	1	86
ARID Black/ NLOS	12	19	12		
WOODED Black/LOS	12	21	12	1	90
WOODED Black/ LOS	12	20	12		
Total - ARID	48	79	48	2	177
Total - WOODED	48	83	48	2	181

OCU/ GUI

Separate monitor for video
Joystick for camera, teleop



Map showing waypoint
2 status logs
GUI for switching to teleop and manipulating camera

Types of Interventions

Type of Intervention	Description	No of Interventions Arid	Wood
Terrain	Interventions due to terrain obstacles (Traction and Tippage)	24	6
Movement	Interventions that affect movement (loopback, snailing, nomotion)	9	4
Plan Failure	Interventions due to a failure in the planner (noplan, deadplanner)	1	3
NeedOperator	Interventions where the Bot asked for operator assistance. This is a level by itself because the cause is not specified and requires operator interpretation to determine correct response.	10	77
Total		44	90

Time for Interventions

	Arid		Wooded	
Type of Intervention	# interv.	Ave. Time (sec)	# interv.	Ave. time (sec)
Terrain	24	154	6	181.5
Movement	9	190	4	102
Plan Failure	1	29	3	229
NeedOperator	10	162	77	136.4
Totals	44	161	90	143.1

SA Time

	Arid		Wooded	
Type of Intervention	Ave SA time (sec)	% interv time	Ave SA time (sec)	% interv time
Terrain	25.1	16.3	17	9.2
Movement	28.4	15.0	4	4.2
Plan Failure	0	0	22	9.8
NeedOperator	31.8	19.5	18	13.3
Totals	26.7	16.6	17.4	12.1

Results

- Time to acquire SA is not insignificant even in ideal conditions (such as this study)
- UI designs have to provide ways for SA to be quickly acquired
  - Manual camera mode was actually faster than automatic in this study
- There is an effect for terrain (or task complexity)
  - Need to develop adaptive user interfaces
- Moreover, how can users “catch up” if they have not been monitoring conditions

Time on Tasks (USAR)

- Analyzed (AAAI 2002 USAR) time spent on tasks
  - Navigation
  - Victim identification
  - Failures
  - Logisitics

			% Time			
Teams	Run	Total Time	Navigation-Monitoring Navigation	Victim ID	Failure	Logistics
A	1	10:39	46	51[a]	0	3
	3	14:45	62	18	19[b]	1[c]
B	1	14:33	81[d]	19	0	0
	3	16:42	77	23	0	0
C	1	13:26	59	23	17[e]	0
	3	14:39	69	12	18[f]	0
D	1	15:12	55	32	0	12[g]
	3	13:30	87	4	0	9

Time on Tasks

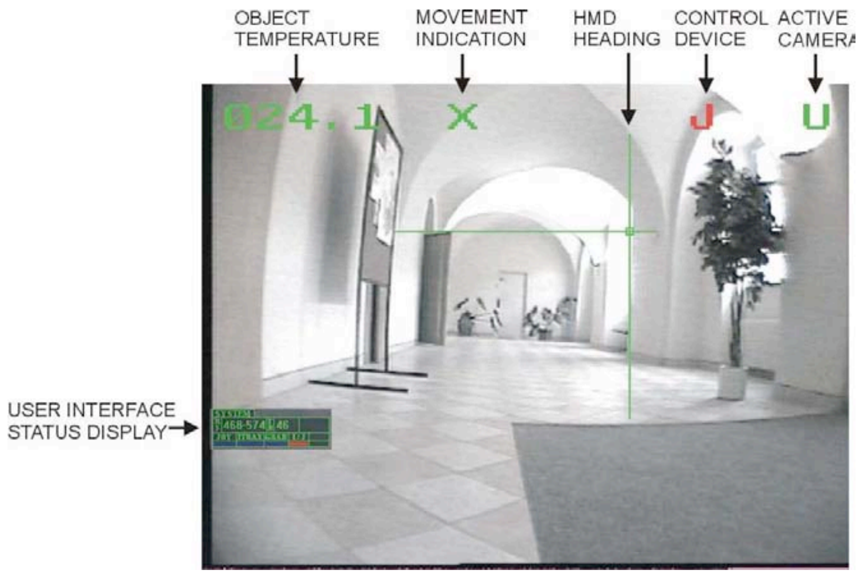
- Additional tasks to consider for USAR
  - Navigation can be broken down into
    - Obstacle extraction
    - Monitoring
    - Commanding (teleoperation, issuing way points, other interaction forms)
  - Mapping activities (human or automatic)
  - User interface overhead (time spent switching between windows, rearranging, time spent switching between cameras, etc.)
- Time spent on various tasks can produce efficiency measures
  - Can correlate with arena coverage, number of victims found

Critical Incident Analysis

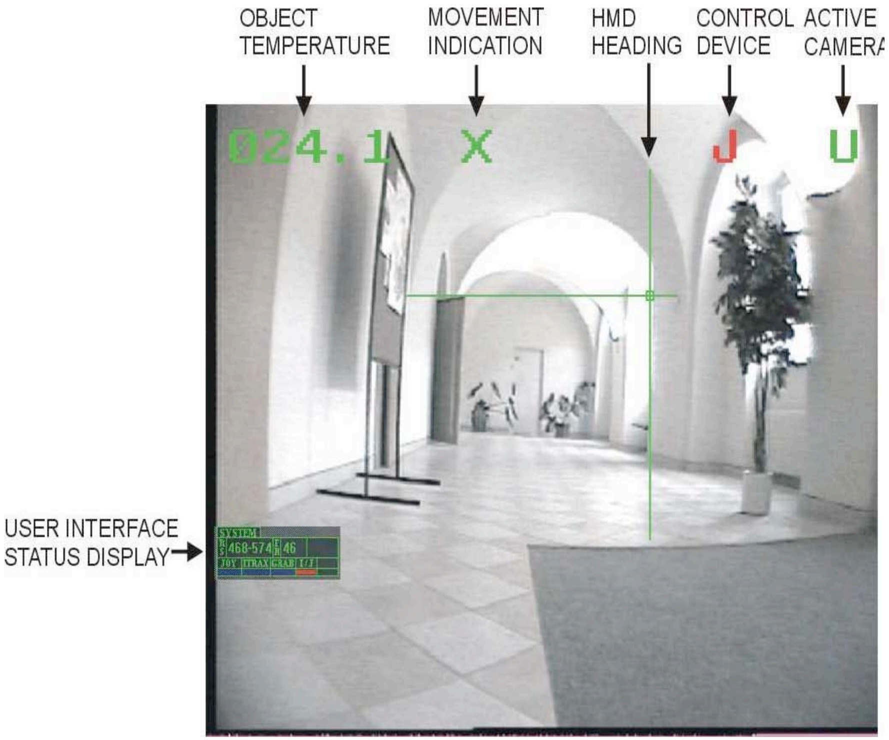
- Applied in the study of USAR competitions in NIST Test Arenas
- Why?
  - Identify potential situations where operator needs to pay attention
  - Identify information that is helpful to operator in dealing with critical incident
- Tasks that were being performed
  - Global navigation
  - Local navigation
  - Obstacle extraction
  - Understanding vehicle state
  - Victim identification
- Measured 'Critical incidents'
  - Situation where the robot could possibly damage itself, the environment or the victim
  - Can have either negative or positive outcome
  - We identify only negative outcomes in this analysis

Methodology

- Analyzed video tapes of robots moving in arena to identify critical incidents
- Three teams who reached the final were analyzed



Team A

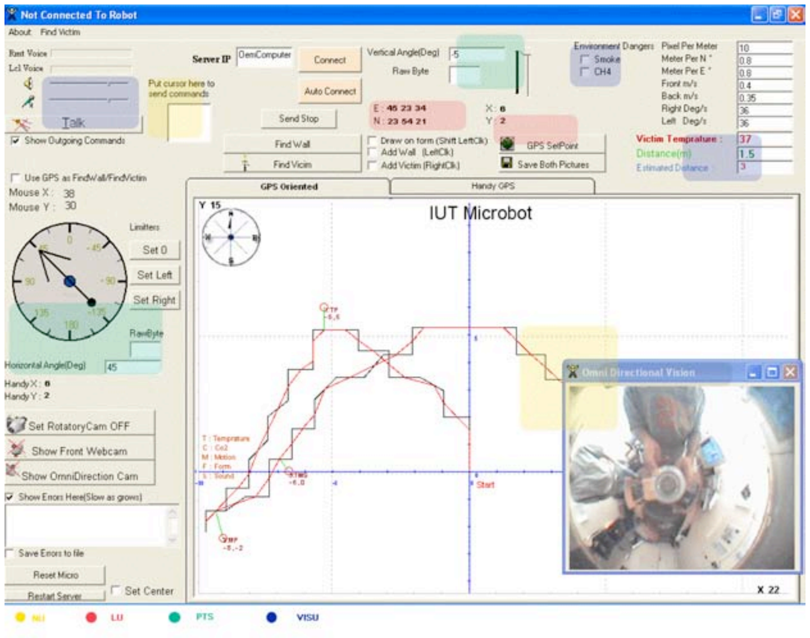


- VR interface
- VR goggles linked to head tracking to move cameras
- Displayed
  - Camera selected
  - Camera position relative to robot body
  - Thermal sensor
  - Audio provided

Team B

- 2 robots, one used only as a communication relay
- 2 video feeds displayed on separate devices
  - Moveable camera
  - Top view of robot + several inches to each side

Team C



- Used two robots but no communications between them
- Displayed (from large robot)
  – Video
  – Omni directional camera
  – Map background used to mark victim location

Methodology

- Two coders coded 12 runs; 2 semifinal runs for each of the 3 teams and 2 final runs
- Two coders coded 8 runs independently and compared
  - 22 incidents located by both coders
  - Kappa coefficient for those 22 incidents was .926
  - 15 incidents that were found only by one or the other coder
  - Agreement on incidents was much easier than finding incidents

Results by Team

Critical Incident	Overall	Team A	Team B	Team C
Global navigation	0	0	0	0
Local navigation	15	4	1	10
Obstacle extraction	26	6	9	11
Vehicle state	12	5	2	5
Victim identification	0	0	0	0

Explanations- Local navigation

Team B <Team A < Team C

- Team A's operator could construct frame of reference using 2 degree of freedom camera; rear cameras also helped
- Team B also used overhead camera to give frame of reference- and operator did not have to remember to use strategy
- Team C had 360 camera but not helpful for local navigation
- Traction is also an issue – most overloaded the interpretation of the video for this

Explanations – Obstacle extraction

Team A < Team B < Team C

- Team A's front and back cameras helped give an understanding of situation.
  - Could move camera to see robot in relationship to obstacle
- Also used audio to hear motors revving

Explanations – Vehicle state

Team B < Team A = Team C

- Team B had less information in GUI than other two teams but robot had impressive mobility.
- Team B had a top down view of robot which may have helped give position information

Exercise (45 minutes)

- Critical Incident Analysis
- Tapes from Robocup 2003
- Teams of 5-6– watch tapes
  - Identify the critical incidents
    - Note times, descriptions
  - Compare with team members
- Can you classify the awareness violations in the critical incidents you identified?
- Given the incidents that you observed, what awareness violation occurred?
  - Local navigation
  - obstacle
  - Vehicle state

Awareness in HRI

- Awareness is used frequently in CSCW
- Definition [Drury 2001]
  - Given two participants p1 and p2 who are collaborating via a synchronous collaborative application...
  - ...awareness is the understanding that p1 has of the
    - presence,
    - identity and
    - activities of p2
- But HRI is different due to
  - Single or multiple humans interacting with a single or multiple robots
  - Non-symmetrical relationships between humans and robots; e.g., differences in
    - Free will
    - Cognition

HRI Awareness Base Case

- Given one human and one robot working on a task together...
- ... HRI awareness is the understanding that the human has of the
  - location,
  - activities,
  - status, and
  - surroundings of the robot; and
- the knowledge that the robot has of
  - the human's commands necessary to direct its activities and
  - the constraints under which it must operate

A General Framework for HRI Awareness

- Given n humans and m robots working together on a synchronous task, HRI awareness consists of five components:
  - Human-robot awareness
  - Human-human awareness
  - Robot-human awareness
  - Robot-robot awareness
  - Humans' overall mission awareness

Details

- Given n humans and m robots working together on a synchronous task, HRI awareness consists of five components:
  - Human-robot: the understanding that the humans have of the locations, identities, activities, status and surroundings of the robots. Further, the understanding of the certainty with which humans know this information.
  - Human-human: the understanding that the humans have of the locations, identities and activities of their fellow human collaborators

Details, Continued

- Robot-human: the robots' knowledge of the humans' commands needed to direct activities and any human-delineated constraints that may require command noncompliance or a modified course of action
- Robot-robot: the knowledge that the robots have of the commands given to them, if any, by other robots, the tactical plans of the other robots, and the robot-to-robot coordination necessary to dynamically reallocate tasks among robots if necessary.
- Humans' overall mission awareness: the humans' understanding of the overall goals of the joint human-robot activities and the measurement of the moment-by-moment progress obtained against the goals.

Examples: Analysis of Awareness Violations

- Analyzed HRI of top four AAAI-2002 teams
- Observed and videotaped robots, interfaces, operators
- Coded activities
- Isolated “critical incidents” and determined causes

Examples of Awareness Violations

- **Lack of human-robot awareness of robots' location**
  - Team A deployed small dog-like robots (Sony AIBOs) off of the back of a larger robot
  - One AIBO fell off and became trapped under fallen Plexiglas but operator didn't know this
- **Lack of human-robot awareness of robots' surroundings**
  - Operator using Team B's robot in "safe" mode became frustrated when robot would not move forward
  - Operator changed to "teleoperate" mode and drove robot into Plexiglas
  - Plexiglas was sensed by sonar and indicated on a sensor map, but map was located on a different screen than video (operator did not check the sensor map)

Examples, continued

- **Lack of human-robot awareness of robots' status**
  - Operator using Team B's robot moved the video camera off center for a victim identification
  - Robot maneuvered itself out of tight area in autonomous mode
  - Upon taking control of robot, operator forgot that camera was still off-center
    - Operator drove robot out of arena and into the crowd

HRI Awareness

- All critical incidents in this analysis were due to a lack of one or more types of awareness
- Human-human and robot-robot awareness criteria may be more useful as scope and complexity of human-robot teams increases

Dialogue Experiment

- Conducted with systems for question-answering in intelligence analysis
- Used a Wizard of Oz study
  - Systems were at different levels of robustness so we used a text chat type of interface over the net
  - The system developers at the other end presented the question to their system
  - They could either use the answer the system returned, modify the answer somewhat, or just answer the question themselves
  - We tracked system, modified, or wizard

Results from Dialogue Studies

- Metrics
  - Number of turns to accomplish a task
    - Number of clarification turns/ total number of turns
  - Ratings from users
    - Naturalness of dialogue
    - Knowing what to say
    - Context
    - System or user in control
    - Terse or verbose
  - Some issues in dialogue
    - Knowing when it is your turn
    - Interrupting
    - Keeping context

Heuristics

- Is necessary information present for human/ robot to intervene?
- Is information presented appropriately?
- Is interaction language efficient for human/robot?
- Are interactions efficient and effective?
- Do interactions scale to multiple robots?
- Do interactions support robot evolution?

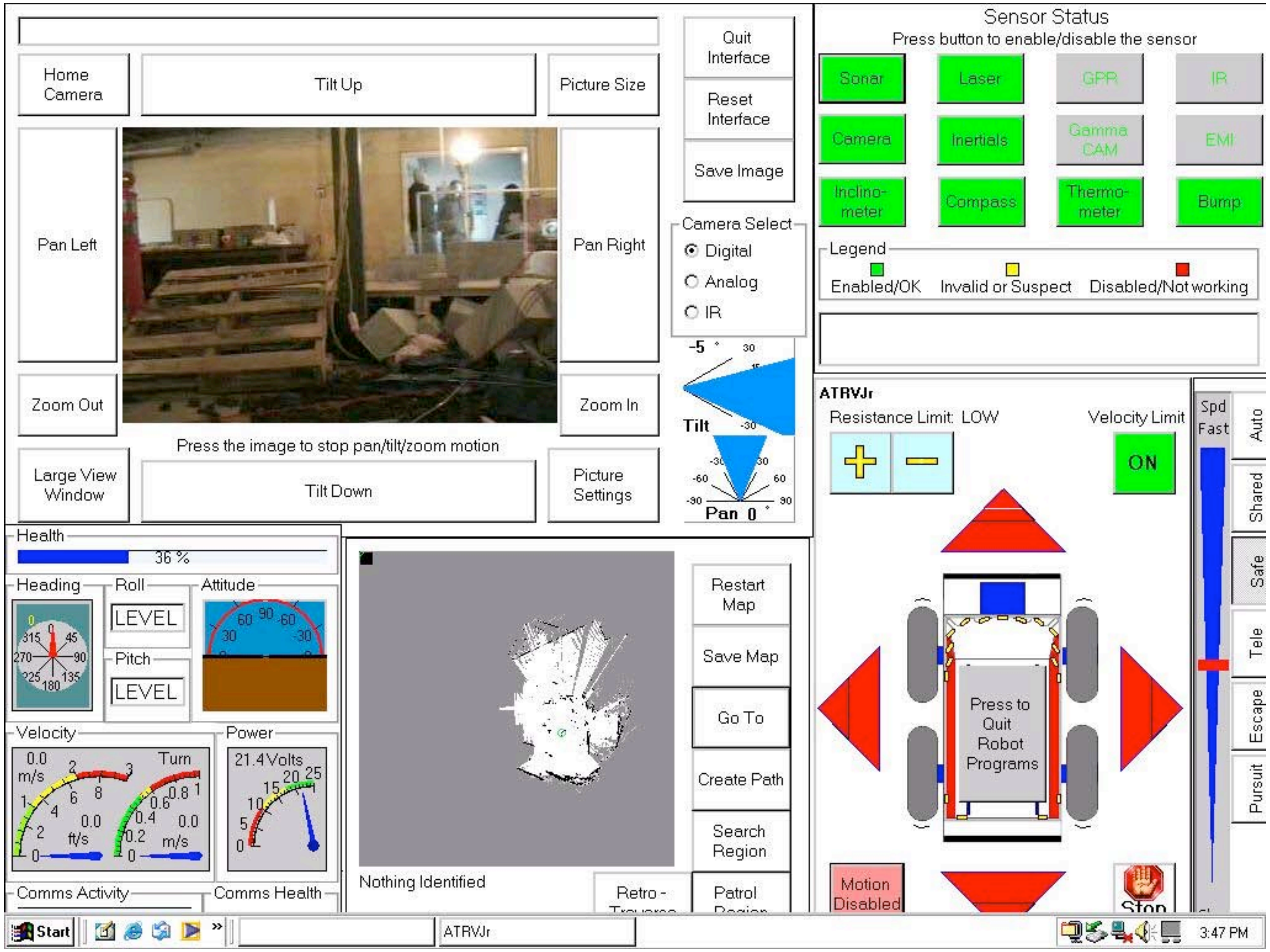
How to Use Heuristics

- Experts in HRI (Usability) walk through a prototype of the user interface
  - Needs to have visual components prototyped but does not need to be fully functional
- Heuristic #1 – needs to be examined for SA information from a domain expert view
- Heuristic #2 – can use some current HCI/ visual presentation guidelines here
- Heuristic #3 – can estimate efficiency of basic tasks, effectiveness looks at possible ways to make errors
- Heuristic #4 – need better design guidelines for multiple robots
- Heuristic #5 – need to determine if more sensors, different autonomy levels can be incorporated

Exercise

- Let’s apply the heuristics

Older Version of INEEL UI

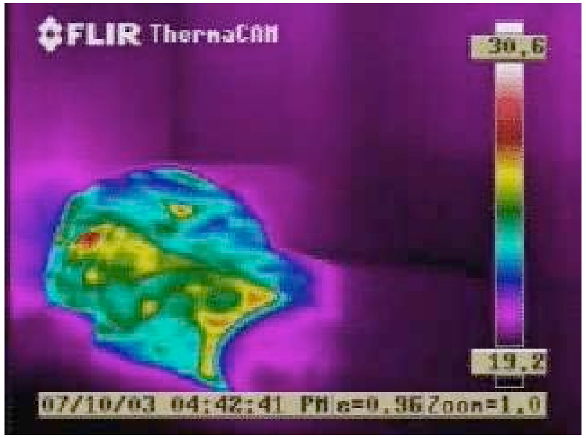


Other Issues with current interfaces/interactions

- Relying solely on video
- Overloading the video
- Duplicate interfaces for multiple robots
  - No notion of “group” control
- Relying on the operator to fuse all information
- No notion of “health” of robot
  - Sensor health; camera position
- Overloading the joystick
  - Camera and drive
- Recall rather than recognition
- Multiple computers for the same interface

An Idea for Combining Sensor Information

Color video image



FLIR image

Combined image:
- Victim was unseen in color video
- FLIR difficult to interpret
- Combined image appears to be looking through the wall

From Hestand and Yanco 2004

Combining Sensor Information

- Display shows combined video, FLIR, sound, and CO2
- CO2 is higher and heat is shown, so there might be a victim here

From Hestand and Yanco 2004

Combining Sensor Information

- Display shows combined video, FLIR, sound, and CO_2
- CO_2 is low and no heat is shown, so there's no victim here
- Plan to experiment with display of additional information and give ability to turn items on or off

From Hestand and Yanco 2004

Guidelines for HRI

from studies conducted at USAR competitions

- Enhance awareness. Provide a map of where the robot has been. Provide more spatial information about the robot in the environment to make operators more aware of their robots' immediate surroundings.
- Lower cognitive load. Provide fused sensor information to avoid making the user fuse the data mentally.
- Increase efficiency. Provide user interfaces that support multiple robots in a single window, if possible. In general, minimize the use of multiple windows.
- Provide help in choosing robot modality. Provide the operator assistance in determining the most appropriate level of robotic autonomy at any given time.

Additional Guidelines for SA (Endsley)

- General
- Certainly design principles
- Complexity design principles
- Alarm design principles
- Automation design principles
- Multioperator design principles

How to develop new guidelines?

- Based on study results
- See www.usability.gov for a good example of usability guidelines for web site
  - These are all rated as to whether they are from
    - Controlled laboratory study
    - Results of a number of usability tests
    - Model results
    - Opinions of experts

www.usability.gov

Guidelines by Chapter

Chapter	Title	Relative Importance	Strength of Evidence	Relative Score
1:1	Set and State Goals	●●●●●	●●○○○	10
1:2	Use an Iterative Design Approach	●●●●●	●●●●●	25
1:3	Evaluate Websites Before and After Making Changes	●●●●●	●●●○○	15
1:4	Provide Useful Content	●●●●●	●●●●●	25
1:5	Understand and Meet Users' Expectations	●●●●○	●●○○○	8
1:6	Establish User Requirements	●●●●○	●●●●○	16
1:7	Use Parallel Design	●●●●○	●●●●○	16
1:8	Consider Many User Interface Issues	●●●●○	●●●○○	12
1:9	Focus on Performance Before Preference	●●●●○	●●●○○	12
1:10	Set Usability Goals	●●●●○	●●●○○	12
1:11	Select the Right Number of Participants	●●●●○	●●●●○	16

Other Possibilities for Areas of Evaluation

- Ubicomp
  - Attention
  - Adoption
  - Trust
  - Conceptual models
  - Interaction
  - Invisibility (inferencing)
  - Impact
  - Appeal

Agenda for Evaluation

- Evaluations and metrics
  - We need to do more evaluations with agreed upon metrics to produce guidelines for HRI
  - Once we know what “good” HRI is we can develop toolkits to help developers
- Challenges
  - Develop evaluation methodologies
  - Determine which metrics distinguish been HRI designs
  - Consider all factors:
    - Expertise of users
    - Domains of use
    - Roles of users
    - Autonomy levels of robots
    - Number and heterogeneity of robots

Research Directions for HRI (Lumelsky and Scholtz)

- Studies of human intervention with different levels of robot autonomy.
- Developing and delivering cues to facilitate remote perception based on cognitive studies.
- Cognitive studies on limitations of human intelligence in typical human-robot tasks (such as limitations in spatial reasoning, reaction speed, consistency, effects of fatigue, etc.)
- Interaction modalities, both input and output, that depart from today's typical means - keyboards, mice, displays - and can be used in various physical environments.
- Appropriate levels of abstraction for effective but intuitive command and control of robots.
- Development of roles for robots and humans within teams, based on studies of human roles, role switching, and handoff behaviors.
- Adaptability of humans, robots, and human-robot teams according to the dynamic nature of situations.
- Scalable user interfaces to allow one human to work efficiently with a team of multiple robots.
- Designing tools for developing human-robot interfaces.
- Robot architectures and world models to support human-in the loop.
- Formation of human-robot teams that support robot evolution.
- Evaluation methodologies and metrics to assess the progress of research in human-robot teams.

Questions?

Backups

Sensors

- People use vision, hearing and other senses to gain knowledge about the state of the world
- Need to understand what we're trying to do in order to determine what we need to sense in the world
- Examples:
  - Trying to cross a street, need to look left and right for oncoming vehicles. Might also listen for cars.
  - Walking in a dark hallway, could use touch to follow the walls
- Robots use sensors to detect conditions in the world

Sensors

- Measure some feature in the world
  - Distance from an object
  - Amount of light
  - Air temperature
  - Loudness of a sound
- Sensors that measure the same feature and process the data in similar ways form a *modality* (refers to raw input)
  - Sound
  - Temperature
  - Light

Types of Sensing

- *Proprioception:* Measurements of movement relative to internal frame of reference
  - Example: number of wheel rotations
- *Exteroception:* Measurements of layout of the environment and objects relative to robot's frame of reference
  - Example: how far away an object is
- *Exprioception:* Measurement of position of the robot body or parts relative to the layout of the environment
  - Example: robot arm bumping into an object

Definitions from *Introduction to AI Robotics* by Robin R. Murphy

Ranging Sensors

- Ultrasonic, or sonar, ranging
  - Sound pulse is emitted
  - Sensor measures how long it takes to hear the echo
  - Distance is computed using known speed of sound
  - Problems: sound pulse can be deflected or absorbed
- Infrared ranging
  - Sensor emits a beam of infrared light
  - Sensor measures how far away an object is by where the reflected beam hits on the sensor or by time taking to see the light's reflection
  - Problems: different types of materials have different reflecting properties
- Laser ranging
  - Sensor emits a laser beam in a horizontal line
  - Height of beam shows how far or close obstacles on a line are
  - Problems: sensor is size of a coffee pot, much more expensive than sonar or infrared ranging

Localization Sensors

- Global Positioning Sensors (GPS)
  - Uses satellite signals to determine latitude, longitude, and altitude
  - Problems: Need to be in line of sight of satellites, so doesn't work indoors
- Dead reckoning
  - Measure wheel rotations to determine how far a robot has traveled and in what direction
  - Problems: Error prone
- Compass
  - Can determine heading of robot
- Tilt sensors
  - Can determine robot's slope in the environment

Cameras

- Video cameras
  - Color or black and white
  - Can send raw video back to the operator or process video on board the robot
  - Problems: Sending back high-quality video can be slow, vision processing is computationally expensive
- FLIR (forward looking infrared) cameras
  - Sense heat in the environment
  - Grayscale or color scale output of heat levels
  - Problems: Can be hard to interpret, expensive
- Stereo vision
  - Provides depth information
  - Problems: Can be expensive to compute

Environmental Sensors

- Temperature
  - Measure the ambient temperature
  - Problems: Not as useful as FLIR
- Gas detectors
  - Can sense different types of gasses (ex. CO2)
  - Problems: Sensors with fast reponse time are expensive, slower sensors could miss detecting the gas if the robot is going too fast
- Microphones
  - Can listen for sounds in the environment
  - Problems: Can be difficult to separate sounds, particularly in a noisy environment

Passive vs. Active Sensors

- Active sensors emit some form of energy into the environment in order to sense the environment
  - Ultrasonic sensors send out a sound wave
  - Laser ranging sensors send out a laser beam
- Passive sensors detect energy levels already in the environment
  - Microphone listens for sounds in the environment
- Typically, active sensors require more power to operate, which can be an issue on a smaller platforms
- In stealth operations, passive sensors are preferred

Levels of Autonomy

Teleoperation
- WTC robots

Shared Control
- Robot wheelchairs
- Mars rovers

Autonomous
- Tour guide robots
- Delivery robots

- Autonomy level is often a continuum, not a single point
- Robots can have adjustable autonomy, sliding up and down the autonomy scale

Teleoperation

- User makes all of the navigation decisions and must monitor the robot’s state at all times
- According to some definition of robots, teleoperated systems are not considered to be robots
- Hard to draw the line between an RC car and a teleoperated robot

Safe Teleoperation

- Sensors can be used to prevent the robot from driving into obstacles
- Operator still makes all of the navigation decisions, but the robot will stop instead of running into obstacles
- Operator must still constantly monitor the robot’s state

Shared Control

- Robot performs some of the tasks
- Operator performs others
- For example, with a robot wheelchair
  - The robot stays centered on the current path and avoids obstacles
  - The rider plans the path to the desired location
- Removes some of the monitoring tasks from the operator, although the operator does need to understand why the robot is taking particular actions

Autonomous Control

- No system is completely autonomous – must be given a goal
- Operator does not need to monitor the robot at all times
- However, monitoring is required if the robot will need the human to intervene at some point
- Difficult to be highly autonomous in unstructured environments

Sliding Scale Autonomy

- Robot systems can be designed with multiple levels of autonomy
- Either the operator or the robot must select the appropriate mode at a given time
- Still an open research area

Presentation of Sensor Information

- Most robot systems have many different types of sensors
  - Sonar
  - Laser ranging
  - Video
  - FLIR
- What's the best way to present this information to the operator?
- How much fusion can occur?