A Practical Approach to Understanding Robot Consciousness

Kristin E. Schaefer¹, Troy Kelley¹, Sean McGhee¹, & Lyle Long²

¹US Army Research Laboratory
²The Pennsylvania State University

Can a robot be conscious?

OR

Is it the degree to which the human interacting with the robot perceives it to be trustworthy?
CONSCIOUSNESS is MORE than just INTELLIGENCE

Can a robot be conscious?

Is CONSCIOUSNESS the same thing as AWARENESS?

- **Subjectivity**: Meaning derived from one's own ideas, moods, and sensations

- **Unity**: All sensor modalities melded into one experience

- **Intentionality**: Experiences have future meaning

- **Others?**

If a robot exhibits emotions, does that make it conscious?

Emotions such as fear, anger, sadness, happiness, disgust, and surprise can be modeled theoretically, and vary due to rewards and punishments.

\[
Emotion(t)_i = w_{o_i} + \sum_{j=1}^{t} \gamma_{i}^{(t-j)} (w_{1,ij}R_{ij}^+ + w_{2,ij}R_{ij}^-)
\]

Eight emotions that vary with time

Fixed coefficients that define temperament

Positive and negative reinforcement

BUT, is trying to design a conscious robot a practical approach?

Example of a MARCbot

Is it the degree to which the human interacting with the robot perceives it to be trustworthy?

Photo: http://www.army.mil/article/19042/Robots_reduce_risks_for_paratroopers/

Photo: Mark Crosby (2015) My Robot Helper
“Never trust anything that can think for itself if you can't see where it keeps its brain.”

~ J. K. Rowling

- Most people do not know how a robot makes decisions.
- So, is it really “seeing” the brain, or our perception of the robot and the actions associated with the decisions that can impact our trust?
Physical Form affects perceptions of trustworthiness

- **Stimuli:** 49 pictures different real-world robots, 7 robot domains
- **Participants:** Over 200 novice participants
- **Findings:** Ratings of perceived intelligence (PI), “robotness” (RC), and negative social influence (SI) can be used to predict trustworthiness of a robot from providing no other information than a picture of the robot.

\[
\hat{Y}_{\text{trustworthiness}} = \text{Constant} + \text{PI} + \text{RC} - \text{SI}
\]

Schaefer, KE (2013) The Perception and Measurement of Human-Robot Trust. Doctoral Thesis. Figure 14

\[ Y_{\text{trustworthiness}} = \text{Constant} + \text{PI} + \text{RC} - \text{SI} \]
Human-Robot Trust

Societal Influence

HRI Information

Pre-Interaction Trust

Measure Attitudes toward robot

Previous Experience

HRI

Post -Interaction Trust

Environment: Team & Task

Robot

Human

Measure Trust Propensity

Measure Information about HRI task

Measure Human States

Expectancy

Measure Initial Trust

Robot Capabilities
LOA
Intelligence
Mode of communication

Perceived Risk

Load Human States

Load Post-Interaction Trust

<<additional interaction with robot>>

(Schaefer, 2013; Figure 14)
“Give me a robot that acts like my bird dog”
~MG William Hix, Deputy Director, ARCIC

A Paradigm shift - from Tool to Team Member

From teleoperation... 
...towards autonomous operation

An Unmanned System that
- Understands its environment
- Conducts useful activity
- Acts independently, but...
- Acts within prescribed bounds
- Learns from experience
- Adapts to dynamic situations
- Possesses a shared mental model
- Communicates naturally
Goal: Move away from over-specialized design to more generalizable decision-making capabilities

A theoretical approach for designing the underlying information processing architecture

Is there a link between the underlying computational architecture and the associated perceptions of the person?

- **Two Algorithms** to identify novel events and enhance episodic indexing

- **Benefits of this approach:**
  - This allows associative cues to be set to novel information
  - Allows the anticipation of future novel events following one exposure to new stimuli

- **New Approach:** This provides the computational justification for episodic indexing of information as a post hoc process
  - Provides justification for certain robot behaviors
Novelty Algorithm

Let $\alpha =$ vector of observations
Let $\beta =$ number of observations in $\alpha$
Let $\mu =$ matrix of observation correlations
Let $T =$ threshold value for a correlation
Let $B =$ % of $T$, give set of observations
Let $\gamma =$ number of correlations that exceed the threshold $T$

\[
\begin{align*}
\gamma & \leftarrow 0 \\
& \text{for } i = 0 \rightarrow \beta - 1 \\
& \quad \text{for } j = 0 \rightarrow \beta - 1 \\
& \quad \quad \text{if } i == j \text{ continue} \\
& \quad \quad \mu_{i,j} \leftarrow \text{correlation}(\alpha_i, \alpha_j) \\
& \quad \quad \text{if } \mu_{i,j} > T \text{ then } \gamma \leftarrow \gamma + 1 \\
& \end{align*}
\]

or, where $x =$ correlation($\alpha_i, \alpha_j$):

\[
\gamma = \sum_{i,j=0}^{\beta-1} f(x) > T \\
\]

Let $\tau =$ % of correlations that exceed the threshold $T$

\[
\tau = \gamma / ((\beta^2 - \beta)/2)
\]

if $\tau > B$ then

RobotStatus $\leftarrow$ Nothing has changed
else

RobotStatus $\leftarrow$ Novel event occurred
end

Episodic indexing allows anticipation of future novel events following one exposure to new stimuli

*Identify the event* prior to the novel event (3e)
Create *New Episode* (nE) starting with the event just prior to the novel event
(nE) = (3e:6e)
Convert *3e to symbolic information* based on current goal (g) and current symbolic perceptual (p) information
(nE) = (e_i (g,p):6e)
Convert *Novel events* (4e:5e) to symbolic (s) event information
(nE) = (e_i (g,p):(s):6e)
Convert 6e to *reinforcement information* (R)
(nE) = (e_i (g,p):(s):R)
Repeat until the end of collected episodes resulting in
\[ \{nE \mid nE = e_n(\text{goal/perception}):(s):R\} \]

Episode (E) = set of events (1e:ne)

Episodic Indexing could help calibrate trust

Do expected behaviors match actual behaviors?
• Improving the underlying architecture could be linked to outward robot behaviors that exude the capability to learn

How does the person know that the robot knows what is going on?
• Appropriate feedback is important to enhancing situation awareness and calibrating trust (Schaefer & Straub, 2016)
• If it is possible to identify early event cues, then it could be possible to provide better feedback timing.

Example: Why did the driverless vehicle stop?

Practical Approach: Near-term robots that can make appropriate decisions in novel, high-risk environments

Successful Human-Robot Interaction: This is based in part on the trust perceptions of the person interacting with the system
- Individuals may have very limited knowledge of how a robot makes decisions or processes information
- All they “know” is based on the behaviors of the robot and the feedback from the robot

Possible Considerations:
- Information processing approach to robot design is fast and relatively simple
- Episodic indexing was found to be efficient process for recognizing novel events and helping to store memories
- Trust Calibration: The concept of episodic indexing could be linked to the timing of robot feedback
Contact Information

Trust: Kristin E Schaefer
kristin.e.schaefer2.ctr@mail.mil

Novelty Algorithm: Sean McGhee
sean.m.mcghee.ctr@mail.mil

Intelligence/Episodic Indexing: Troy Kelley
troy.d.kelley6.civ@mail.mil

Information Processing/Robot Emotions: Lyle Long
lnl@psu.edu