Measuring Cognitive Workload:
From HMI Assessment to Real-Time State Monitoring

Bruce Mehler – Research Scientist | MIT AgeLab & New England University Transportation Center (bmehler@mit.edu)

Car HMI Conference, Detroit
April 23, 2018

(This deck supplements material that was presented during the talk along with additional reference details.)
Multi-disciplinary team at MIT studying aspects of the evolving advanced automotive technology ecosystem and it’s impact on the future of mobility

- **Advanced Vehicle Technology (AVT) Consortium**
  - Studying currently available automated vehicle technologies
  - Implications for future technology development, policy, insurance, acceptance, etc.

- **Advanced Human Factors Evaluator for Automotive Demand (AHEAD) Consortium**
  - Driver workload metrics for HMI evaluation and driver monitoring
  - Situational awareness, attention management in an increasingly automated ecosystem

- **Driver workload evaluation**
  - Basic research on physiology, eye movements, etc.
  - Applied system evaluation

- **ADAS and highly automated vehicles**
  - Human centered considerations in trust, attention, situational awareness, etc.
  - Factors that may influence the adoption of automated vehicle systems
  - Non-verbal communication between road users

- **Applications of Deep Learning**
  - Driver monitoring
  - External scene perception
Why Do We Care About Cognitive Workload?

• If engaging with a user interface is mentally demanding, we are not going to have a happy customer...

• If physical controls are difficult to manipulate or visual content of a HMI are difficult to understand
  o a user is likely to spend more time looking off road
  o Safety becomes a concern

• For HMIs that don’t necessarily require a user to look off the road – there is still the threat of cognitive absorption
  o slower responses
  o or even “look but do not see”
How Do We Measure Workload?

• Variety of Methods (each with advantages & disadvantages)
  – Self-report
  – Impact on performance (driving and/or secondary task)
  – Physiological arousal

• Performance and physiological measures offer the advantage of being **objective** and relatively **continuous**

• Theoretical advantage of physiological measures
  – Investment of effort to maintain performance may result in little change in performance measures at lower levels of load
  – But, be detectable as increased physiological activation.
What Is Practical to Measure in the Car?

Initial evaluation started in 2006 in a simulator and involved 204 healthy young adults in a series pilot and formal studies. Follow-on work has involved 600+ participants of various ages and health status driving instrumented vehicles on real highways, as well as hundreds of participants in subsequent applied simulator based research.

Measures considered:

• Heart Rate (& HRV)
• Pulse height (PPG - peripheral blood flow)
• Skin Temperature
• Electrodermal Activity (GSR, SCL)
• Respiration Rate & Depth
• Muscle Tension (EMG)
• Grip Pressure (steering wheel)
• Pupil Diameter
• EEG (brain waves)
• fNIRS (brain blood flow / oxygenation)
• Stress Hormones

Overall question - which measures will prove practical & most sensitive at differentiating levels of demand?
Experimental Manipulation of Stress / Cognitive Demand

A delayed-digit recall working memory - cognitive workload reference task (n-back)

• Series of single digit numbers (0-9) presented in random order aurally at 2.25 sec intervals

• Subject instructed to respond with nth digit back

• Secondary cognitive demand is manipulated by the n-level

• Across levels
  o Auditory demands constant
  o Vocal demands “relatively” constant

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>6 9 1 7 0 8 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blank-back</td>
<td>- - - - - - -</td>
</tr>
<tr>
<td>0-back</td>
<td>6 9 1 7 0 8</td>
</tr>
<tr>
<td>Response</td>
<td></td>
</tr>
<tr>
<td>1-back</td>
<td>- 6 9 1 7 0</td>
</tr>
<tr>
<td>Response</td>
<td></td>
</tr>
<tr>
<td>2-back</td>
<td>- - 6 9 1 7</td>
</tr>
<tr>
<td>Response</td>
<td></td>
</tr>
</tbody>
</table>
Classical Physiological Metrics Are Sensitive to Graded Levels of Cognitive Demand

Demand from n-back delayed digit recall working memory task resulted in:

- Heart rate (HR) changes essentially linear with demand; rapid recovery
- Skin Conductance (SCL) reactivity at low demand suggests emotional component; slower recovery

These findings validate these metrics as continuous measures of cognitive load that are clearly useful for research and evaluation purposes.

Question – if auditory vocal tasks can reliably be used to induce cognitive load, how much cognitive load is present when interacting with voice-based interfaces when driving?
Bars represent mean visual demand in terms of total glance time eyes are off the forward roadway (TEORT). Standard error shown as horizontal line. Dots show the 85% point in the sample distribution for each task.

• Physiological measures and self-report in our initial studies suggested only moderate levels of cog load for voice-based tasks.

• Voice-based radio tuning is less visually demanding than a classic visual-manual HMI,

• but it is not visual demand free, and

• TEORT for newer tasks such as voice-based destination entry can involve significant visual engagement over the duration of the task.
TEORT Address Entry & POI Across Multiple Voice-Based Multi-Modal HMIs

But, is a simple metric such as TEORT the last word on demand in modern HMIs? (Perhaps not.)

The solid horizontal line represents the NHTSA TEORT threshold. The dot above each bar represents the 85% point in the sample distribution for each task. Bars represent individual address entry tasks (e.g. 1st bar is 177 Massachusetts Ave. Cambridge, 2nd bar is 293 Beacon Street Boston etc.)
Broadening scientifically valid perspectives and methodologies for the **objective measurement of demand** placed on drivers by in-vehicle systems and technologies during vehicle use, while considering the increasing role of attention support and management.

As a way of moving beyond previous efforts, AHEAD is considering:

- The role of spatial and temporal characteristics of a task
- A framework in which demand can be optimized across all dimensions, i.e. visual, auditory, haptic, vocal, manual, etc., by taking into consideration the relative cost and benefit interactions of various input, output and processing modalities
- Interactions between secondary tasks and the broader operating environment

We aim to move the language of assessment from **one focused on distraction**, to one that emphasizes **driver attention management and safe operation**, such that **demands on driver, active safety systems, and other higher order forms of automation** can be considered as a whole.
Off-Road vs. On-Road Glance Duration & Crash Risk

Reanalysis of the 100-car naturalistic driving dataset comparing mean single glance duration for:

- off-road glances (left) and on-road glances (right), and
- crash (black) and non-crash events

(Error bars represent 95% CIs.)

Measuring Attention over Time

For modeling visual attention, the concept of an ‘Attention Buffer’ can be used to describe visual allocation and it’s relative impact on driver situation awareness:

- Driver needed to look back at the road and attend to it in order to refresh their awareness of the road situation.

- Driver’s focus away from road - awareness of road situation declining

- Depth of buffer reflects sampling rate required to keep vehicle in lane (every 1.8-2s; cf. Senders, 1967) – potential changes with automation, etc.

Figure adapted from Kircher & Ahlström (2009)
Hybrid Measures of Attention Provide a More Comprehensive View of Safety

**A larger loss of SA preceded crash** epochs in the 100 Car-Data

**Texting Crash** epochs in the SHRP 2 NDS data start with depleted awareness


Important Take-Away

Glance Patterns Tell a More Complete Story of the Effect of HMI Demands on Driver Attention

• Standard off-road glance measures like mean single glance duration and TEORT calculations do not fully capture the effects of HMIs on how drivers are managing their attention over time and space.

• Hybrid metrics like AttenD provide a way to measure the combined spatial and temporal effects of glance patterns on a driver’s stored knowledge of the scene when interacting with HMIs.

• The AHEAD effort has been actively refining the buffer construct specifically from the perspective of understanding attention management & support
  
  o Refining decrement and increment rates based on experimental evaluation and analysis of naturalistic datasets
  
  o Considering the distribution of glances ON road reflecting cog load
But What About Measuring Cognitive Workload When the Driver is Looking on the Road?

- But what about the “thousand yard stare”?
- How do I measure cognitive load when drivers are looking at the road?
- Do I need to measured heart rate, etc. in addition to eye behavior?

Our early work using traditional eye tracking looked at standard deviation of horizontal gaze (plot to right) and found that it was sensitive to detecting initial increases in cognitive load (baseline driving to 0-back to 1-back), but could not differentiate medium from high cognitive load. Thus, peripheral physiological measures (heart rate & skin conductance) have proven superior for that purpose for research measurement.

But what about advances in deep learning techniques for potentially picking out changes in demand associated with more subtle eye behavior changes? The results are shown on the next slide.
Applying A New AgeLab Approach to Cognitive Load Estimation

Using deep learning to learn from and predict the level of cognitive load from the same dataset used in the earlier peripheral physiology studies:

• **Input:** Sequence of raw images of eye region from a single camera
  - 6 second windows (15 fps / 90 images)

• **Model:** Three Dimensional Convolutional Neural Network

• **Output:** 86.1% average accuracy of 3-class prediction
Prediction of cognitive load from micro-position & micro-movements of the eye from a single monocular camera

The video to the right shows an implementation of cognitive load estimation at three levels (low, medium, and high) running dynamically to predict the state of drivers during natural driving.

- This is a less hardware intensive and non-invasive sensing approach that provides sensitivity across the same range of demand levels that we see with medical grade heart rate and skin conductance measures.

- We are working to integrate cognitive load measurement with our modified visual attention buffer model.

The levels are scaled against labeled load states from our original on-road data collection of drivers doing the multiple levels of the n-back task.

Emotional State Estimation in the Context of Driving (& HMI Evaluation)

Which of these drivers is satisfied with the performance of their voice interface system?
Emotional State

Classifying driver frustration through combined video and audio analytics.

Video of driver showing frequent and long duration off-road glances while interacting with a multimodal voice-based interface to compose and send texts. Driver commented that she probably should not be texting but it is ok because of being a voice system.

Video removed for participant confidentiality.
**MIT has established an Advanced Vehicle Technology (AVT) consortium to combine resources to study how real drivers do or do not interact with ADAS & other autonomous driving technologies in the real world.**

- Combined study of users in their own cars (1 year+) and MIT vehicles (1 month).
- Approximately 1000 miles of video, audio, GPS, accelerometer and CAN data is being added to the dataset per day.
- The Cadillac CT6 is now being added to the research fleet.
- AVT members include OEMs, suppliers, insurance companies & others. New members are welcome.

---

**MIT Autonomous Vehicle Technology Study**

<table>
<thead>
<tr>
<th>Car Model</th>
<th>Miles</th>
<th>Days in Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla Model S</td>
<td>14,117 miles</td>
<td>248 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>10,271 miles</td>
<td>366 days</td>
</tr>
<tr>
<td>Tesla Model S</td>
<td>5,186 miles</td>
<td>91 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>3,710 miles</td>
<td>133 days</td>
</tr>
<tr>
<td>Tesla Model S</td>
<td>24,657 miles</td>
<td>588 days</td>
</tr>
<tr>
<td>Tesla Model S</td>
<td>16,666 miles</td>
<td>353 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>15,074 miles</td>
<td>276 days</td>
</tr>
<tr>
<td>Tesla Model S</td>
<td>9,188 miles</td>
<td>183 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>5,111 miles</td>
<td>232 days</td>
</tr>
<tr>
<td>Tesla Model S</td>
<td>3,005 miles</td>
<td>144 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>3,005 miles</td>
<td>144 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>4,596 miles</td>
<td>132 days</td>
</tr>
<tr>
<td>Tesla Model S</td>
<td>1,306 miles</td>
<td>69 days</td>
</tr>
<tr>
<td>Tesla Model X</td>
<td>1,306 miles</td>
<td>69 days</td>
</tr>
</tbody>
</table>

Study months to-date: 21  
Participant days: 7,146  
Drivers: 78  
Vehicles: 25  
Miles driven: 275,589  
Video frames: 3.48 billion  
Study data collection is ongoing.  
Appendix: Selected References


Appendix: References on MIT Work with Production Voice Systems


Appendix: References on MIT Work with Production Voice Systems (cont.)


Discussion / Questions

For Further Follow-up:

Bruce Mehler bmehler@mit.edu