How to Engineer Trust into AI-based Systems – an Automotive Perspective

M. Siegel – OFFIS e.V
Building a better world ..... 

> Smart health and well-being, clean energy, sustainable transportation, climate action, responsible production, sustainable cities, inclusive societies, ....
> Cf. Grand Societal Challenges in Europe, Sustainable Development Goals of the UN, ...
> (less philanthropic: generating higher company profits)

..... through new technologies, digitization, smart systems
Game Changer AI - Mastering the (yet?) unknown ..... 

1-2 days energy forecast for balancing energy market

1-15 min energy forecast for net stability

Sufficiently good models available -> “conventional” system engineering

no (good) models available but lot’s of data -> learning / AI

- Approximation of otherwise (yet) unknown function
- Black box – inner working “untransparent”
Example : AI-based perception for automated driving

Consensus: AI required for SAE level 3-5 driving

Expert statements: AI will allow to reduce car accidents by >90% in the next 10 years

“No one really knows how the most advanced algorithms do what they do.”
MIT Review, Will Knight, April 11, 2017
Concerns affecting AI-Acceptance

Personal profiling towards surveillance?
Orwell, Huxley, Eggers, ....
social scoring, loss of freedom, ...

Superintelligence?
Alpha Go zero, Skynet, ......
Self programming systems, loss of human control, blind reliance on technology, loss of jobs, ....

Underlying ethical / societal norms and values?
Health treatment decisions, credit worthiness, discriminating legal case assessments, ..... 

Trust in Safety?
Tesla, Uber accidents
Mal-functions of e.g. automated car endangering my life or others .....
Engineering trust into safety-critical AI systems

Automotive R&D directions

Turning „black box“ into „grey/white box“
<table>
<thead>
<tr>
<th>Automotive R&amp;D Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficient, structured Data Acquisition &amp; Data Qualification</td>
</tr>
<tr>
<td>Knowledge Explication &amp; Extraction</td>
</tr>
<tr>
<td>Design-for-verifiability – Architectures and Algorithms</td>
</tr>
<tr>
<td>Robustness of Decisions against noise, manipulation</td>
</tr>
<tr>
<td>Self-explainability capabilities for human-machine-cooperation</td>
</tr>
<tr>
<td>Completeness of acquired knowledge and self-awareness of knowledge gaps</td>
</tr>
<tr>
<td>Safe, incremental knowledge adaptation and updates</td>
</tr>
<tr>
<td>Verification &amp; Validation of AI-components / systems</td>
</tr>
<tr>
<td>Safe-guarding of application / knowledge boundaries depending on ASIL levels</td>
</tr>
</tbody>
</table>
Supporting R&D Activities

Pre-competitive cooperation:
R&D roadmaps, joint projects, platform standardization, ....

Shared Data
Data interoperability, joint trainings data, incidents databases, continuous update processes, ....
## Interdisciplinary Research on social acceptability of AI

### Interdisciplinary Research – Philosophy, Legal, Social Science, Psychology,....

<table>
<thead>
<tr>
<th>Building value systems into AI-based decision making</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principles of justifiable decisions:</td>
</tr>
<tr>
<td>• individually desirable</td>
</tr>
<tr>
<td>• societally beneficial</td>
</tr>
<tr>
<td>• democratically controllable</td>
</tr>
<tr>
<td>• legally and ethically justifiable</td>
</tr>
<tr>
<td>Genuine user dialog: self-explanations, generation of justifications, debating mechanisms between system/human, transparent decision-making processes</td>
</tr>
<tr>
<td>Compliance of emerging AI-behavior with norms and principles of our society</td>
</tr>
</tbody>
</table>

......
Today’s system complexities go beyond the “comfort zones” of engineers and product owners – especially in shrinking time-to-market windows.

Not only AI systems but also many of today’s complex control systems are susceptible to potentially safety-critical mal-function during operation.

Paradigm shift: whole system engineering is changing towards virtual validation / certification, fast safe system-updates along whole product lifecycle and systematic in-field learning (DevOps).
Thank You

Questions?