The topics discussed during the clinic

https://hait.cs.lth.se/_media/emergence.jpeg

https://hait.cs.lth.se/_media/implementation_methods_for_industry.jpeg

https://hait.cs.lth.se/_media/silent_failures.jpeg

https://hait.cs.lth.se/_media/trust_and_transparency.jpeg

Eight topics were suggested and discussed in the workshop (in no particular order):

- Cognitive offload
- Hybrid cognitive systems
- Shared workload
- Emergence
- Implementation methods for industry
- Individualised AI
- Silent failures
- Trust and transparency

The following paragraphs try to summarise the most important aspects that can be distilled from the notes. Passages, phrases, concepts, or simply words that have been filled through educated guessing by the editor (Elin), are marked in brackets […]. Terms that are taken directly from the notes are set in *italics*.

Topics or terms that received markers are coded in colour as follows:

1 marker

2 markers

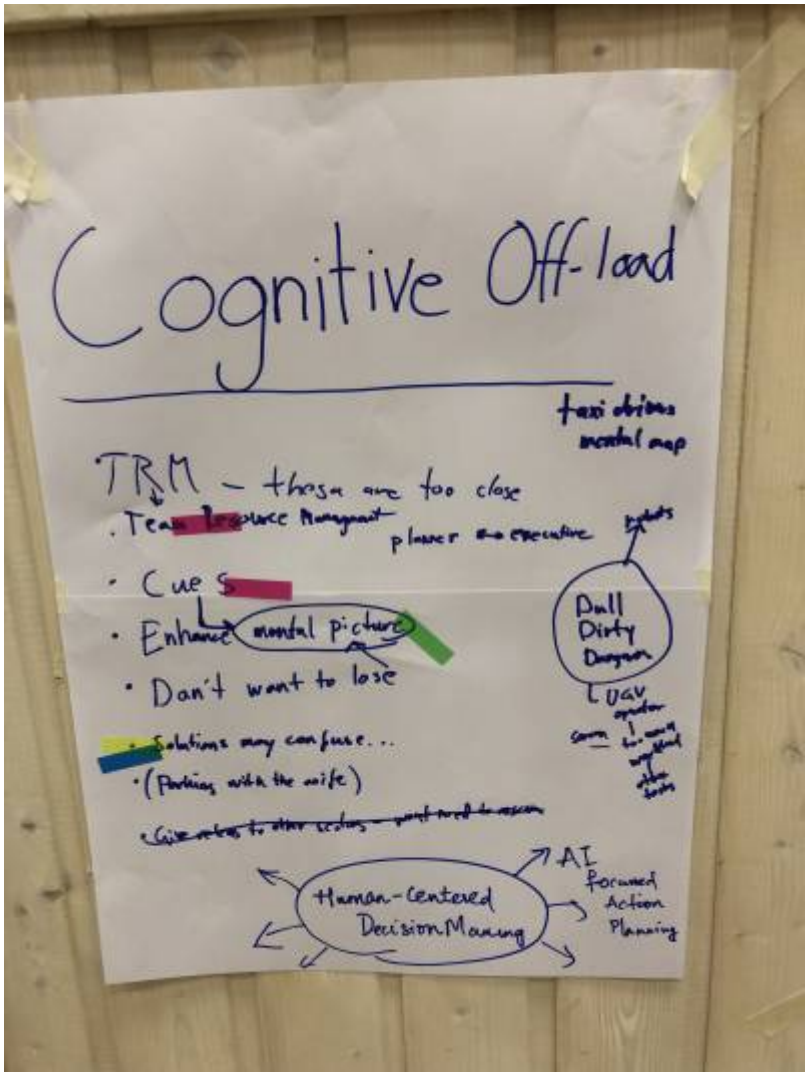
3 markers

4 markers

Cognitive offload

Aspects mentioned and marked as important in this discussion were *Team Resource Management* (TRM) as well as the idea of *providing / enhancing cues* for a *mental picture*, [to give operators overview], as the aim would be to not loose the *mental picture* of the situation at hand. However, [suggested] *solutions may confuse* [solutions to providing overview, or solutions to directly solving a problem?]. Other items mentioned in this discussion were *human-centered decision making*, with one particular area *AI focused action planning*. *Dull, dirty and dangerous tasks* link *robots* / machines with operators (example UGV operators), where the issues of generally too *high workload* and handling *other tasks* at the same time come into play.

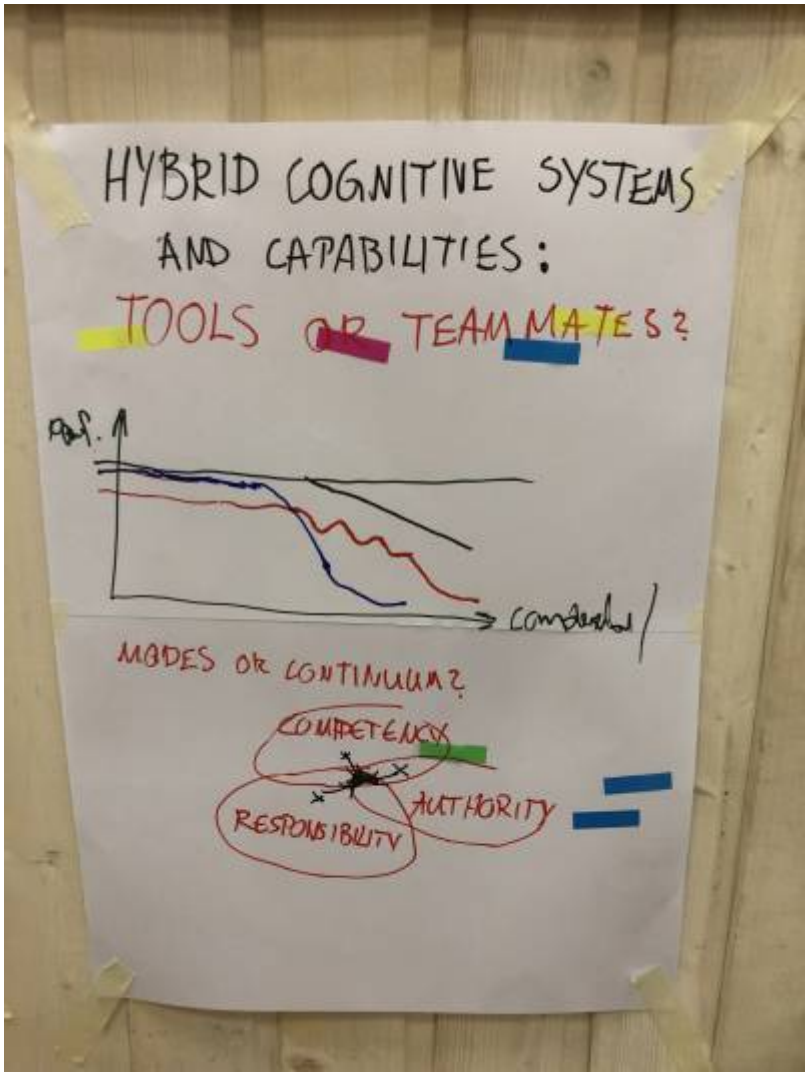
See also the original chart from the workshop:



Hybrid cognitive systems and capabilities

The main questions discussed here were whether such systems should be seen / promoted as *tools or teammates* and whether a transition between these viewpoints should be seen as a *mode switch or a continuous (gradual) transition*, which could then also be defined in *three dimensions (competency, authority, and responsibility)*.

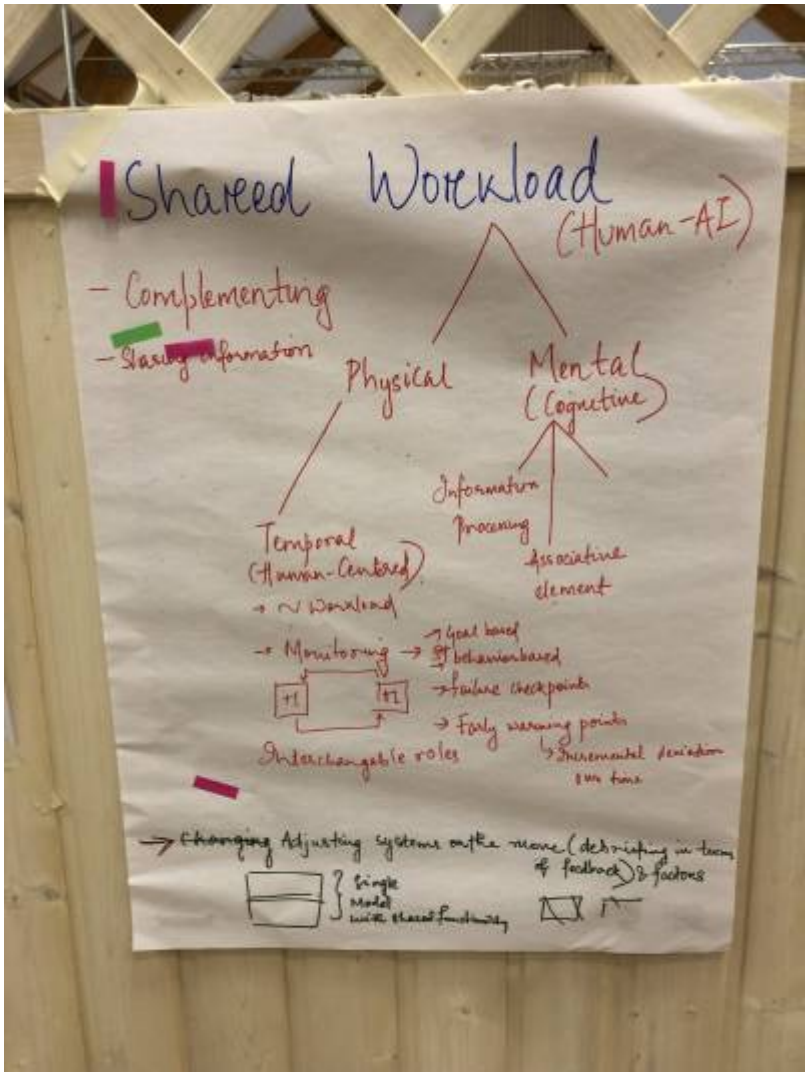
See also the original chart including two diagrams from the workshop:



Shared workload

Regarding the topic of *shared workload*, a taxonomy was suggested, considering two types of workload to be discussed, namely *physical* and *mental / cognitive* workload. On the physical side, temporal aspects of sharing workload and in particular monitoring these efforts were raised. Monitoring of a jointly handled task can be done goal based or behaviour based, and there could be failure checkpoints and early warning points [signals] considered. One further question would be to discuss who is monitoring whom and in how far these roles would be interchangeable. On the mental side, aspects as *information processing* and *associative elements* were shown as sub-categories. Aside the taxonomy, an important aspect discussed was the overall objective of sharing workload, i.e. to allow for human and AI to *complement* each other and to *share information*. It was also mentioned out that system design plays a role, where it might be valuable to look into *systems that can be adjusted "on the move"* [to find the right way of sharing the workload for the task at hand].

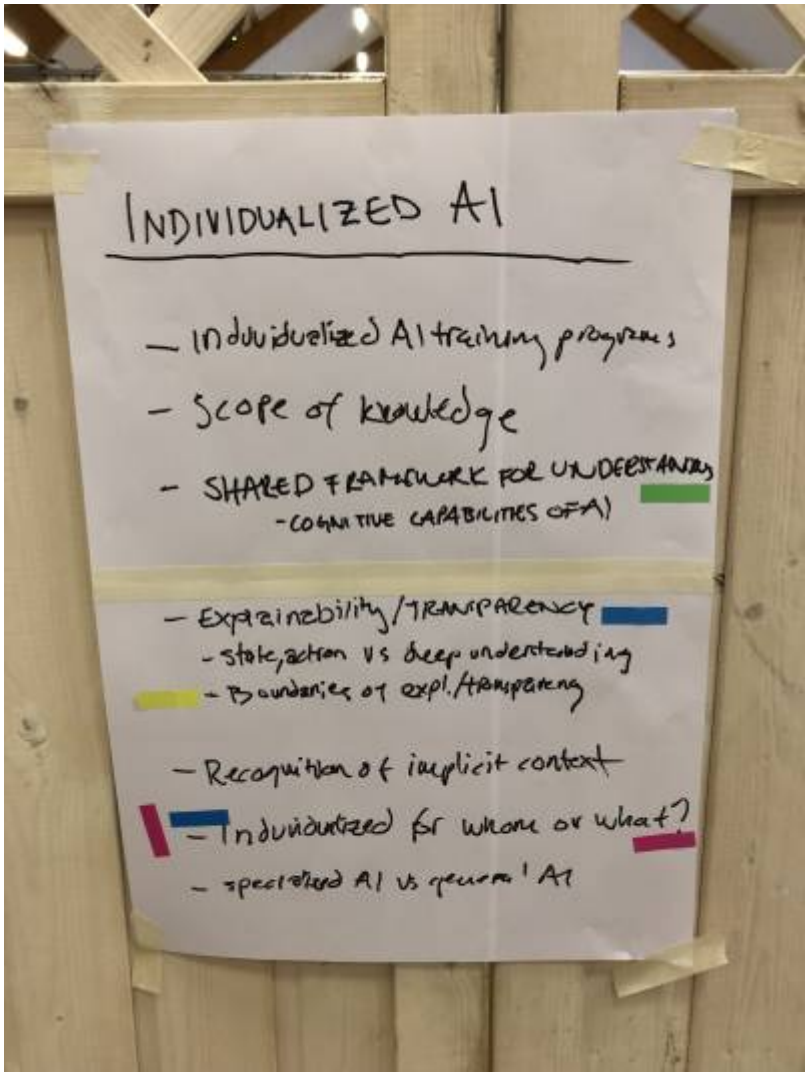
See also the original chart from the workshop:



Individualised AI

The discussion here centered around the question whether (and if so, how) AI should be individualised, both regarding its application and the training necessary to apply it, as expressed in the call for *individualised AI training programs*. Aspects like the *scope of knowledge* and a *shared framework for understanding* [somewhat limited by the cognitive capabilities of AI]. One point mentioned here was *explainability / transparency*, and in particular the degree to which this should be provided (on *state / action level or to provide deep understanding*) , i.e. the *boundaries of explainability and transparency* came into play here. This included also the question of recognition of *implicit context*, [that might be different given different individuals being part of the team]. A central question is that of *whom or what [an AI / system] should be individualised* for, and an overarching question is that of [whether individualisation should be discussed in terms of] *general or specialised AI*.

See also the original chart from the workshop:



Trust and transparency

Two aspects, that would link to the following topic (How to deal with silent failures), are that [the AI] *must understand why it made a mistake* and to *explain when disagreeing*. For trust to be established, [the system / human?] *must be predictable*, which can include *simulation* as a tool. This altogether is, however, not *necessarily inherent to AI based systems, it could also be a "normal" algorithm* that is discussed. Further, the issue of *liability* is mentioned. Topics for research are suggested as *visualising AI decision making (also for different types of users* and *digital twin*.

See also the original chart from the workshop: