A References Architecture for Human Cyber Physical Systems, Part II: Fundamental Design Principles for Human-CPS Interaction

KLAUS BENGLER, Technische Universität München
WERNER DAMM, Carl von Ossietzky Universität Oldenburg, Germany
ANDREAS LUEDTKE, DLR - Institut für Systems Engineering für Zukünftige Mobilität, Oldenburg
REIGER JOCHEM, Carl von Ossietzky Universität Oldenburg
BENEDIKT AUSTEL, DLR - Institut für Systems Engineering für Zukünftige Mobilität, Oldenburg
BIANCA BIEBL, Technische Universität München
MARTIN FRÄNZLE, Carl von Ossietzky Universität Oldenburg, Germany
WILLEM HAGEMANN and MORITZ HELD, Carl von Ossietzky Universität Oldenburg
DAVID HESS, Vanderbilt University
KLAS IHME, DLR - Institute of Transportation Systems, Braunschweig
SEVERIN KACIANKA, Technische Universität München
ALYSSA J. KERSCHER and LAINE FORREST, Vanderbilt University
SEBASTIAN LEHNHOFF, Carl von Ossietzky Universität Oldenburg
ALEXANDER PRETSCHNER, Technische Universität München
ASTRID RAKOW, Carl von Ossietzky Universität Oldenburg
DANIEL SONNTAG, Carl von Ossietzky Universität Oldenburg und DFKI-Deutsches Forschungszentrum für Künstliche Intelligenz, Nds.
JANOS SZTIPANOVITS, Vanderbilt University
MAIKE SCHWAMMBERGER, MARK SCHWEDA, and ANIRUDH UNNI, Carl von Ossietzky Universität Oldenburg
ERIC VEITH, OFFIS e. V., Oldenburg

As automation increases qualitatively and quantitatively in safety-critical human cyber-physical systems, it is becoming more and more challenging to increase the probability or ensure that human operators still perceive key artifacts and comprehend their roles in the system. In the companion paper, we proposed an abstract reference architecture capable of expressing all classes of system-level interactions in human cyber-physical systems. Here we demonstrate how this reference architecture supports the analysis of levels of communication between agents and helps to identify the potential for misunderstandings and misconceptions. We then develop a metamodel for safe human machine interaction. Therefore, we ask what type of information exchange must be supported on what level so that humans and systems can cooperate as a team, what is the criticality of exchanged information, what are timing requirements for such interactions, and how can we communicate highly critical information in a limited time frame in spite of the many sources of a distorted perception. We highlight shared stumbling blocks and illustrate shared design principles, which rest on established ontologies specific to particular application classes. In order to overcome the partial opacity of internal states of agents, we anticipate a key role of virtual twins of both human and technical cooperation partners for designing a suitable communication.

CCS Concepts: • **Human-centered computing** → *Interaction paradigms*; **HCI theory, concepts and models**;

Additional Key Words and Phrases: Real-time systems, Cyber-Physical Systems, architecture, interaction design, Human-CPS Interaction

ACM Reference format:

This work is supported, in part, by the United States National Science Foundation Office of International Science and Engineering (OISE) PIRE program and the Directorate of Computer and Information Science and Engineering (CISE) CPS program under grant OISE-1743772, and in part by the German Research Foundation (DFG) under grants for projects "Assuring Individual, Social, and Cultural Embeddedness of Autonomous Cyber-Physical Systems," project numbers 433524510, 433524788, and 433524434.

Authors' addresses: K. Bengler and B. Biebl, Technische Universitaet Muenchen, Boltzmannstr. 15, 85748 Garching b. Muenchen/Germany; e-mails: bengler@tum.de, bianca.biebl@tum.de; W. Damm, J. Reiger, M. Fränzle, W. Hagemann, M. Held, S. Lehnhoff, A. Rakow, M. Schwammberger, M. Schweda, and A. Unni, Carl von Ossietzky Universität Oldenburg, Ammerlaender Heerstrasse 114-118, 26129 Oldenburg/Germany; e-mails: werner.damm@uni-oldenburg.de, jochem.rieger@uni-oldenburg.de, martin.fraenzle@uni-oldenburg.de, willem.hagemann@uni-oldenburg.de, moritz.held@uni-oldenburg.de, sebastian.lehnhoff@uni-oldenburg.de, a.rakow@uni-oldenburg.de, m.schwammberger@uni-oldenburg.de, mark.schweda@uni-oldenburg.de, anirudh.unni@uni-oldenburg.de; A. Luedtke and B. Austel, DLR - Institut Systems Engineering für zukuenftige Mobilitaet (SE), Escherweg 2, 26121 Oldenburg/Germany; e-mails: andreas.luedtke@dlr.de, benedikt.austel@dlr.de; D. Hess, Department of Sociology, Vanderbilt University, PMB 351811, Nashville, TN 37235-1811 USA; e-mail: david.j.hess@Vanderbilt.Edu; K. Ihme, Deutsches Zentrum fuer Luft- und Raumfahrt, Institute of Transportation Systems, Lilienthalplatz 7, 38108 Braunschweig/Germany; e-mail: klas.ihme@dlr.de; S. Kacianka and A. Pretschner, Technische Universität Muenchen, Boltzmannstr. 3, 85748 Garching b. Muenchen/Germany; e-mails: kacianka@in.tum.de, alexander.pretschner@tum.de; A. J. Kerscher and F. Laine, Vanderbilt University, Nashville, TN 37240, USA; e-mail: alyssa.j.kerscher@vanderbilt.edu, forrest.laine@vanderbilt.edu; D. Sonntag, Carl von Ossietzky Universität Oldenburg, Ammerlaender Heerstrasse 114-118, 26129 Oldenburg/Germany and Deutsches Forschungszentrum für Kuenstliche Intelligenz GmbH (DFKI), Interaktives Maschinelles Lernen, Marie-Curie-Strasse 1, 26129 Oldenburg/Germany; e-mail: daniel.sonntag@uni-oldenburg.de; J. Sztipanovits, Institute for Software Integrated Systems, Vanderbilt University, PMB 351829, Nashville, TN 37235-1829 USA; e-mail: janos.sztipanovits@vanderbilt.edu; E. Veith, OFFIS e. V., Escherweg 2, 26121 Oldenburg/Germany; e-mail: eric.veith@offis.de.

2378-962X/2024/01-ART3
https://doi.org/10.1145/3622880

Klaus Bengler, Werner Damm, Andreas Luedtke, Reiger Jochem, Benedikt Austel, Bianca Biebl, Martin Fränzle, Willem Hagemann, Moritz Held, David Hess, Klas Ihme, Severin Kacianka, Alyssa J. Kerscher, Laine Forrest, Sebastian Lehnhoff, Alexander Pretschner, Astrid Rakow, Daniel Sonntag, Janos Sztipanovits, Maike Schwammberger, Mark Schweda, Anirudh Unni, and Eric Veith. 2024. A References Architecture for Human Cyber Physical Systems, Part II: Fundamental Design Principles for Human-CPS Interaction. *ACM Trans. Cyber-Phys. Syst.* 8, 1, Article 3 (January 2024), 27 pages.
https://doi.org/10.1145/3622880

1 INTRODUCTION

The companion paper [14] introduced a reference architecture for designing human cyber-physical systems. Although it highlights the type of information to be perceived by the ego system on each of its layers, thus describing information flow on a *logical level*, it abstracts completely from the challenges in realizing this information flow. This article focuses on the challenges and principles of *human perception and cognition*: given the increasing levels of automation, how can we increase the probability that those artifacts of the environment of a human ego system, which are key for correct control decisions at any of its internal layers, can be presented to the human under real-time constraints in a way, that he or she directs his or her attention to exactly these critical artefacts? How can we validate that the sensory input provided is actually perceived and interpreted without misunderstandings?

In the first section of the article, we use a well-documented case study on the interaction in an aircraft cockpit [16] to exemplify how instantiating the reference architecture for this application helps to highlight the large space of possible misconceptions and misunderstandings that could potentially endanger the safety of an aircraft. In this example, missing the need for an additional unexpected confirmation to enter the next layer in controlled airspace is only avoided by the redundancy in the cockpit, where the second officer—only through gestures—alerts the first officer that an additional radio communication is required.

The cockpit example shows how inter-cockpit communication fundamentally relies on a shared "cockpit speak," which is anchored in a domain ontology, including not only all relevant artifacts but also procedures: without these, timely interaction in the cockpit would be impossible. But even more important is the cockpit "speak"—i.e., the agreed-upon representation of artifacts of the ontology, be it through a dedicated cockpit jargon of English, through gestures, through gazes, or through visual or acoustic feedback of instruments in the cockpit.

The second section of this article presents fundamental principles shared across application classes, where professional or at least semi-professional trained humans interact—potentially as a team—with cyber-physical systems to control a shared environment. We highlight shared stumbling blocks and illustrate shared design principles, which fundamentally rest on established ontologies. We anticipate a key role of virtual twins (c.f. Section 2.2 in [14]) of both human and technical cooperation partners as well as of the controlled system and its environment as an enabler for designing a suitable representation of the most urgent and/or relevant information to be exchanged.

Our deliberate focus on human ego systems reflects the overall focus of this series of papers on human cyber-physical systems. The challenges in guaranteeing confidence in the perception/cognition chain of highly autonomous technical ego systems are currently addressed in a number of projects. They are also addressed in the list of references in the companion paper [14].

2 A CASE STUDY IN THE INSTANTIATION OF THE REFERENCE ARCHITECTURE FOR HUMAN CYBER-PHYSICAL SYSTEMS: DISTRIBUTED COGNITION IN AN AIRLINE COCKPIT

The case study used to instantiate the reference architecture is described by Edwin Hutchins in his article "Distributed Cognition in an Airline Cockpit" [20]. The scenario he proposes utilizes a high-fidelity flight simulation of a Boeing 727-200 with a crew of three pilots: the captain, first officer, and second officer. The captain acts as the "pilot not flying"; his role is to oversee communications and monitor the two "pilots flying" (the first and second officers in charge of operating the aircraft).

At the time of analysis, the plane is in route from Sacramento to Los Angeles and climbing through 19,000 feet toward a cruise altitude of 33,000 ft. The following transcript is contrived from audio and visual recording observations and details the minute-and-a-half exchange between the pilots and **air traffic control (ATC)** during flight operations.

Time	Speaker	Message
0216	S/O	Xxx nasa nine hundred
0224	S/O	Departure report
	S/O	Nasa nine hundred from eh Sacramento to Los Angeles international we have eh /. . . / fuel on board twenty-seven point eight fuel boarded is not available out time is one six four five up time is one six five five

Hutchins's analysis of the crew reveals shared cognition as a system-level property where the distributed access to information among actors and background ethnographic understanding are the basis for a functional system, independent of individual operator shortcomings. He proposes a concept of intersubjectivity that allows humans to derive meanings in non-verbal behaviors and to assess the crew's expectations, violations, and actions and how they pertain to the distribution of information storage. He follows the propagation and movement of information throughout the system, asserting the existence of two pathways for information processing that can occur from either the formulation of expectations or the development of proofs to substantiate observations.

The system cognitive properties that he highlights are:

- individual knowledge,
- physical property representation,
- the organization of representations,
- distributional characteristics of knowledge, and
- access to task-relevant information across crew members.

A consideration of these properties is significant for the holistic development and application of an accurate metamodel of anthropocentric systems. The following adaptation of the reference architecture presented in the companion paper [14] to the case proposed by Hutchins attempts to explicate the beliefs, perceptions, and maneuvers of each of the four agents of the cockpit system (the captain, first officer, second officer, and autopilot-system) to better understand functional system operations and assess shortcomings in cockpit interactions.

2.1 The Pilots

We choose to analyze this scenario from the perspective of the cockpit starting with the human actors. They share many qualities such as their environment and training background, but they must be differentiated with regard to their roles.

Normative Environmental Layers

The applied reference architecture is described from the perspective of the singular ego

0247	Capt	Oakland center nasa nine hundred request higher {First officer reaches to vicinity of altitude alert setting knob when ATC begins transmission}
0254	OAK24L	Nasa nine hundred /…/ roger contact oakland center one thirty two point eight {First officer pulls his hand back from the altitude alert knob when ATC says "contact Oakland Center." 2.5 seconds after the end of ATC transmission. First officer looks at captain.} {Captain looks at first officer}
0300	F/O	Thirty two eight
	Capt	Thirty two eight?
	F/O	yeah
	Capt	ok
0303	S/O	That's correct, nasa nine hundred
	Capt	\one three two eight ah, nasa nine hundred {Captain twists knob on radio console} {First officer looks in direction of captain}
0315	Capt	Center nasa nine hundred twenty one point seven for two three zero requesting higher
0323		{Second officer turns toward front of cockpit}
0325		{First officer looks at captain}
0325	OAK15H	Nasa nine hundred /…/ oakland center climb and maintain flight level three three zero and expedite your climb please
0327		{First officer reaches the altitude alert as ATC says "expedite your climb" S/O turns to the performance tables on the second officerwork surface.}
0331	F/O	ok
0333	Capt	Three three zero nasa nine hundred {Captain leans toward and looks at first officer} I didn't catch the last part
0336	F/O	Expedite your climb
	Capt	ok
0339		{Second officer reaches thrust levers and pushes them forward}
0341	Capt	That's firewall thrust {Captain looks at first officer}
	All	(laugh)

system, so relevant conceptions of the normative environment are expressed in the ENV1 and ENV2 layers.

ENV1: Accepted ethical, societal, and psychological principles

The first environmental layer includes accepted ethical, societal, and psychological principles such as aviation ethnography, implicit moral and institutional codes of conduct through which the pilots operate, and human psychological phenomena of social cognition and interaction. The semantics, customs, and habits that develop as a result of years of aviation experience lead to a unique shared understanding of the pilots. Hutchins terms this their ethnography, which can be seen by the esoteric discourse that he calls "aviationese" riddled throughout the script. Had the reader no background knowledge of aviation training, a true understanding of the script would be rendered impossible. It is through this ethnographic knowledge and language that system actors formulate expec-

tations and derive meaning from their perceptions. For example, in the case study, the captain violates the first officer's ethnographically based expectation to conduct a readback of the instructed frequency change. There is a formulaic structure of radio talk that is formally accepted within the community.

The accepted ethical principles in this case can be most explicitly understood through the pilots' agreement to a code of conduct as outlined by licensure requirements and airline contractual obligation. ENV1 also encompasses the more implicit learned, developed, and innate ethics in which the human actors (and by extension the technical systems that were produced by these actors) operate.

Finally, the psychological principles acknowledged in this layer come from theories of behavioral and cognitive conditions. The most relevant example Hutchins has identified is the Speech Act Theory [20]. According to this idea, there are three aspects of speech that a speaker utilizes to convey a message: locutionary (what a speaker actually says), illocutionary (the force of what is said), and perlocutionary (the intended effect of what is said). An understanding of this principle is necessary to interpret the information exchanges between the captain and first officer. Through a mere perlocutionary inquisitive glance, the first officer communicates his expectation of a readback that the captain meets with a confused look. The first officer then interprets this puzzled facial expression and tells the captain, "Thirty two eight." This locutionary utterance holds the abbreviated indirect meaning of "The frequency that you missed is one hundred thirty-two point eight." Due to the socialization and training of the pilots, communication is succinct and occasionally incomplete.

Of course, cultural relativism necessitates that the understanding of ENV1 must be based on the context of the ego, which may exist in cultures with different moral frameworks. For the purpose of this sample, the culture is vaguely classified as the Western aviation community.

ENV2: Applicable regulation, laws, and physical laws

In ENV2, the applicable regulation, laws, and physical laws are here identified as governmental aviation administration laws, company-specific procedures and regulations, and physical performance criteria. In this study, both the **Federal Aviation Administration (FAA)** regulations and the airline's policies apply. Within the case study, company procedures require that the altitude alerter be set whenever a new altitude is assigned to the aircraft as a cautionary form of physically stored memory. For additional formalized prudence, flight condition reports are company required. The second officer occupies himself with filling out these forms during the first portion of the flight.

For each aircraft, there are required operation manuals that outline physical limitations and performance capabilities due to physical law constraints. These manuals include empirically derived figures as well as thorough descriptions of technical system operations. When the crew is instructed to expedite the climb in the case study, the second officer references aircraft manuals to determine the appropriate **Engine Pressure Ratio (EPR)**. The manuals define the maximum operational thrust that will not compromise the structural integrity of the aircraft.

Actual flight-time operations rely on physical law abidance. Most basically, these include the four forces that act upon a plane: thrust, drag, lift, and gravity. The geometry of the aircraft—namely the wings—impacts this interaction, as does the weather, speed, and atmospheric conditions. Climate figures are typically a compilation of recorded measurements from aircraft instruments as well as external weather service information data reports such as the **Automated Terminal Information Service (ATIS), Automated**

Weather Observing System (AWOS), Terminal Aerodrome Forecasts (TAFs), and **Meteorological Terminal Air Report (METAR).** Cockpit instruments are calculated, calibrated, and developed based on observed physical phenomena. A most basic example would be the essential pitot static and gyroscopic instruments that are based on fluid dynamic principles.

For this study, the physical laws pertaining to changing elevation are most relevant. Pressure differences due to these changing elevations can alter airplane performance. Corresponding assumptions are made; 29.92 in. of mercury as an altimeter setting indicates pressure altitudes at or above 18,000 ft MSL (5,486 m). Additionally, basic phenomena such as maximum thrust limitations and radio proximities resulting in a frequency change from Oakland Departure to Oakland Center apply to this environmental layer.

Ego System Levels

A Note on Expectations and Beliefs

The shared aviation ethnography is the basis for the existence of pilot expectations and the subsequent formulation of pilot hypotheses, which are then tested to form beliefs with a sufficient confidence level. Beliefs are based on the pilot's perceptions, which can potentially be selective and primed based on a heavy reliance on past experiences of experts. As humans learn, they make shortcuts and use heuristics with reduced sets of information. Attention and information processing is refined and focused in order to optimize results despite severe processing capacity limitations of human brains (see Section 3 for an extensive discussion of these). As a result, over time, beliefs can form that are limited to perceptions that humans strongly expect. This is where the ENV1 and ENV2 types of principles and regulations are supposed to add a safety layer.

Information processing (i.e., cognition) in humans is closely integrated with expectations, which can compensate for the severe processing capacity limitations. Biases introduced by expectations can prime information processing already at the level of perception (see Section 3). Although an expert's heuristics can increase system efficiency, in particular in systems with limited processing capacity like human brains, they can also result in a potential vulnerability despite ENV1 and ENV2 safety layers. Thus, consideration of information not included in the heuristics is of importance at the reflection layer.

A Note on Roles and Actors

Anchored in this application of the reference architecture are distinctions between the roles of each subsystem. It is not always the case that roles are explicitly associated with each actor. Instead, the roles are often dynamic, and they can entail different, case-dependent responsibilities. Roles influence the expectations formed by system actors based on ethnographic groundings, history, memory, and understandings of environmental laws and phenomena. Furthermore, a pilot fitting other actors into schemata associated with their respective roles can lead to the formulation of second-order beliefs of the crew partners, as provided by the Reference Architecture (see Section 3.2 of [14]). Moreover, hierarchies that are too strong can lead to the dismissal of reflections on dysfunctional situations, like communication glitches.

Within this case study, the pilots share the high-level plan of operating the aircraft to a cruising altitude of 33,000 ft MSL. For the purposes of this illustration, we determine this to be the relevant overarching objective for the pilots within this case study. Strategies and maneuvers are role dependent and are outlined within this case.

EGO1: Value System Level

The value system level for the Pilot Ego includes relevant regulations and operational procedures governing technical and interpersonal interactions among pilots. These include the pilot's internalized understanding of company-specific processes and regula-

tions, FAA policies and regulations, and relevant ethnographic groundings that give rise to operational procedures and interpersonal interactions. An example is the structured group of radio responses that are customary within the aviation community and are manifestations of pertinent societal principles to which the pilots subscribe. Another example of FAA policies is the ATC clearance required when operating in controlled airspaces. The class of these airspaces requires different regulations; for instance, the Sacramento International Airport and Metro Oakland Airport are both class C. The expectations and beliefs that the pilots formed based on these relevant ethnographic groundings give rise to their operational procedures. At the value system level, the maintenance of the safety and comfort of the aircraft passengers, energy optimization, and other overarching values fall under this category.

EGO2: Reflection Level

This is a crucial layer for safety because the human actors will continuously reflect on the plausibility of their beliefs and attempt to reconcile them with any higher-level system violations (c.f. Section 3.3 in [14]). For the pilots, these assessments are conducted through the comparison of representational states across a series of media. Speech, memory, expectations, and physical memory lead to meaningful conceptualizations of state representations that are then compared to validate beliefs.

Physical memory, as Hutchins defines it [20], is based on physical states that are used as a form of memory. Examples of concrete representations of beliefs include instrument readings, written notes, and manual and automatic instrument inputs. These stored representations act as another form of redundant information storage to combat local failures. The Altitude Alert System in this case operates as such a mode.

ATC readbacks for error checking are another case example of the reflection level. In the case scenario, the first officer confirms his frequency reading to the captain as a form of error checking. Additional reflection is conducted by the first officer, leading to the correction of the captain despite the fact that listening to ATC was not the first officer's role. The example points to the concept of intersubjectivity and shared cognition of the pilots within the cockpit system. Additionally, consistency in ethnographically derived expectations among the pilots is ensured at the reflection level.

The next three layers contain the specific thought processes of the pilots during the case study scenario. We will have a closer look at their objectives, expectations, beliefs, and actions as can be inferred from the transcript.

EGO3: High-level Planning

Objective: The aircraft should climb to a cruising altitude of 33,000 ft MSL.

Expectations: ATC will grant clearance, and other pilots will act according to their roles.

Beliefs: High confidence perception of the environment leads to beliefs about the situational setting, consisting of attentive observations, the other pilots, and the instruments. This confidence is formed from the pilot's attentive observations of relevant instruments and actors.

EGO4: Strategy Level

To achieve their common high-level plan, pilots have different strategies according to their different roles in the cockpit.

Captain

Expectations: Must act as "pilot not flying," who contacts ATC and oversees operations of the first and second officers. Strategies: Plans to contact ATC on a given communication frequency.

Beliefs: Captain believes that the radio is accurately tuned based on visually derived understandings reinforced by redundant actors and expectations.

First Officer

Expectations: Must act as "pilot in command" for this portion of the flight. Plane will operate as usual. Other pilots will act according to their roles to execute the shared objective.

Beliefs: First officer believes that operations are consistent with expectations based on situational perceptions.

Strategies: Plans to oversee the maintenance of desired heading and ensure standard safe aircraft operations. Additionally, plans to adjust the altitude alert knob setting upon confirmation from ATC.

Second Officer

Expectations: Must act as "pilot second in command."

Beliefs: Instrument readings are accurate, and flight management input is correct as well.

Captain	**First Officer**	**Second Officer**
Expectations: ATC will grant clearance upon initial contact. *Beliefs*: Pressing and holding the radio transmission button at given frequency calls ATC receiver. *Maneuver*: Calls ATC at given frequency to request for clearance. The request is not immediately granted, violating the expectation.	*Expectations*: ATC will grant clearance upon initial contact. The request is denied, creating new beliefs. (On the reflex level) First officer reaches for altitude alert knob in anticipation. His beliefs are inconsistent with his expectations, so he withdraws his hand. *Beliefs*: ATC did not grant clearance but instead gave a frequency change of 132.8 Hz. *Maneuvers*: First officer does not alter the instruments.	*Expectations*: ATC will grant clearance, and second officer operations are not immediately required to fly the plane. Believes that figures he is inputting within the flight management plan and corresponding paperwork are correct. Expects a syntax or customary methodology of paperwork based on learned experience in the field. *Beliefs*: Figures he perceives from displays and inputs onto paperwork are accurate. *Maneuvers*: Works on required airline paperwork.
Frequency change that Oakland Center indicates is 132.8 Hz.		
Expectations: Center gave a command and a new frequency was stated. *Belief*: Current frequency input is insufficient. Captain hears new frequency, but his belief is not confident enough to execute a maneuver upon until confirmed by the first officer, who then confirms the captain's perception. This confirmation acts as a form of redundancy, resulting in a belief formation of sufficient operable high confidence. *Maneuver*: Captain changes frequency to 132.8 Hz.	*Expectations*: Expecting a readback from the captain to ATC confirming the frequency change. *Beliefs*: Based on the perception, first officer believes (with low confidence) that the pilot heard the command. First officer tries to reinforce this belief to ensure high confidence but contextually concludes that captain is not confident of the instructed frequency. First officer interprets a pause and confused expression from the captain. *Maneuvers*: Looks at captain imploringly, verbally confirms frequency of 132.8 Hz. This is not an explicit role of the first officer at this time, but shared cognition and distributed access to information allows for error checking.	*Expectations*: Climb will be steady and standard. This is violated by ATC's command to expedite, and a new expectation surrounding a hastened climb is adopted. *Beliefs*: Reported figures in the aircraft manual are accurate, as are airspeed indications, wind data, and other environmental quantitative identifiers. *Maneuvers*: Second officer refers to engine pressure ratio (EPR) performance tables to adjust thrust levers accordingly as instructed by Oakland Center. These tables indicate a maximum permissible thrust given the air temperature and altitude at that given time. This action is also an example of a lower layer taking precedence over the value system (in this case violating energy efficiency). Extreme situations can override the value system based on lower layer beliefs.
Expectations: The new frequency will grant permission to climb to the desired cruising altitude. *Belief*: When the radio transmission button is pressed, Oakland Center Departure is contacted on this new frequency. *Maneuver*: Calls new frequency ATC to ask for clearance again. Request is conditionally granted, and the aircraft must expedite the climb.	*Expectations*: Listens for phrase "climb and maintain," which is common in the aviation community and signifies clearance. *Beliefs*: Clearance is granted and altitude knob alerter is now incorrect and must be changed. *Maneuvers*: Sets the altitude alert knob to 33,000 ft MSL.	

Captain	First Officer	Second Officer
Expectations: Clearance was granted with an additional command from ATC. *Belief*: Captain heard musings of an additional phrase after clearance was granted. Once again, belief is of operable high confidence when the first officer reads back a confirmation. *Maneuver*: Confirms to ATC that aircraft has received command through a readback of aircraft identification number (NASA 900). Captain asks first officer what additional information he missed. First officer clarifies the instruction to expedite the climb to the captain.	*Expectations*: All members of the cockpit hear the command to expedite the climb and will act accordingly. *Beliefs*: Aircraft is climbing too slowly. Captain did not hear the final phrase. First officer has a belief of high confidence that he heard accurately. *Maneuvers*: Increases pitch of the aircraft.	
Expectations: The pilots in command will act according to their roles to expedite the climb. The conversation with ATC has concluded. *Belief*: Aircraft and team are operating according to expectations. *Maneuver*: Verbally confirms shared understanding of instruction. Illocutionary act of saying "That's firewall thrust!"		

Strategies: Will be responsible for manipulating thrust and taking care of required paperwork formalizations and other housekeeping tasks.

EGO5: Maneuver Level

The maneuver level is where actions are executed by the pilots. Here, we go through the maneuvers of the three pilots in chronological order. Each row depicts the valuations of concepts of the reference architecture at one time frame. By analyzing the three pilots side by side, we can illustrate the difficulties of collaborating entities.

EGO6: Reflex Level

The reflex level for the Pilot Ego includes trained behavior that does not require attentive cognition. For instance, the habitual, structured calls and responses that the captain conducts would be considered reflexive, as well as the first officer controlling the yoke and the second officer controlling the thrust levers in this case study. The most evident reflex layer control from this example occurs when the first officer reaches for the thrust lever expecting immediate permission to climb but has to withdraw his hand when the ATC response is not congruent with his expectations.

EGO7: Integrity and Health State Management

The seventh ego layer of the integrity and health state management of the pilot system includes pilots trained to compensate for errors of themselves or their crew. This is evident in the case study when the first officer is listening to ATC for frequency assurance when the captain is not confident that he heard it correctly. Though this is technically out of the scope of the explicit role of first officer, the dynamic nature of his position allows for an added degree of safety. Additionally, the pilot's maintenance of sufficient personal physical and mental health is necessary for optimum system function. This is ensured through health checks, breaks, multiple members switching roles during a long shift, and other prudent personal safety measures.

2.2 Ground-truth Ego Environment

Ego(ENV) refers to the ground truth of all relevant artifacts of the ego system's environment as seen by an omniscient observer. In this case study, the ego's environment includes relevant environment plans, strategies, capabilities, and dynamics. Only relevant environment systems deemed of notable risk are outlined. Also, this analysis is narrowed to be within the physical scope of the operating Oakland airspace.

ENV3, ENV4, ENV5

For the first three environment levels, we identify three relevant entities to consider: the autopilot, the ATC (Oakland Center), and the other aircraft in the airspace.

	Autopilot	**ATC Communicator**	**Other Aircraft in Airspace**
ENV3	Objective: Climb to inputted altitude (33,000 ft MSL) while maintaining inputted heading.	Objective: Maneuver aircraft in airspace.	Objective: Reach their destinations.
ENV4	Fly according to a given heading and climb at a steady rate.	Plan flight paths based on radar-generated trajectories. Contact each aircraft in the airspace to grant clearance.	Contact ATC for clearance. Fly on a heading directed toward the destination.
ENV5	Disengage when manually overridden and re-engage when instructed.	Interpret computer-generated data. Contact each aircraft within airspace via a single frequency with instructions to maneuver if necessary. Grant clearance.	Await a break in communication channel to be granted this clearance. Insert maneuver to fly.

Because of a shared set of operational rules, it can be assumed that constituent elements of the shared superstructure (see Section 3.1 of [14]) are acting predictably and obey the rules and regulations of ENV1 and ENV2. The environment plans, strategies, capabilities, and dynamics are assumed to be under control of the shared superstructure of the ATC airspace, complying to an agreed-upon dynamic. The three identified entities of *ego(ENV)* from the perspective of the Pilot Ego are the Autopilot system, Oakland Center, and the other aircraft in the airspace. Their respective strategies and capabilities are listed in the table above.

ENV6: Environment Classification, State, and Belief

The classification of surrounding systems for the pilots is determined through the geospatial orientation parameters:

- Weather
- Time of day
- Cockpit interior (technical systems, other pilots)
- Oakland airspace
- An altitude of 19,000 ft MSL and increasing

ENV7: Uncontrollable EGO State, Capabilities, and Dynamics

The following principles are identified to be the most prominent contributors to EGO7:

- Any health hazards or physical limitations of pilots (such as heart attack, stroke, and visual impairments)
- The failure of any of the mechanical, hydraulic, or electronic subsystems within the cockpit

2.3 The Autopilot Ego

To demonstrate the application of the metamodel to cyber-physical systems, we create an instantiation of the autopilot system analogous to the pilots. Autopilot engagement is assumed for the entirety of this case scenario.

Normative Environmental Layers

ENV1: Accepted ethical, societal, and psychological principles are assumed to be entered at the time of the system's design. Consequently, we assume that regulatory rules and similar design constraints were entered and abided by.

ENV2: Applicable regulator rules, laws, and physical laws can be found in the operation manual for the aircraft and include physical limitations, performance capabilities, and details regarding the performance of its technical systems and software. The design parameters are assumed to follow regulation adherence and operate within physical laws. Finally, the same physical laws as for the Pilot Ego are applied to the Autopilot Ego.

Ego System Level

EGO1: Value System is assumed to be based on values entered at design time as parameters, or values can be automatically adapted based on the current state of the system. The design values are expected to be adapted to the designer's values and to be commensurate with the pilots' value system. Examples include maintenance of safety, energy optimization, prescription of company regulations, and federal policies.

EGO2: Reflection Level refers to the autopilot comparing data from redundant instruments and sensors.

EGO3: High-level Planning deals with the objective to climb to the inputted altitude while maintaining flight management system parameters. This high-level planning relies on the belief that the inputted data is correct and reliable, as revealed by the redundant instruments.

EGO4: Strategy Level involves the strategy to achieve the plan, which is to fly according to the given heading and to climb at a steady rate. The required beliefs are that the physical properties are as inputted to the system from the pilots or otherwise recorded from accessory instruments and sensors (altitude, speed, etc.).

EGO5: Maneuver Level holds the belief that the plane is in a steady, energy-efficient rate of climb. Some new beliefs are introduced with overridden inputs from the pilots, for example, an increased rate of climb following ATC's queue of other flights. The new thrust setting and overridden first officer control of the yoke violate the former belief. The performed maneuver alters the performance when the presets are manually overwritten by the second officer's manipulation of the throttle and re-engaging with the new settings.

EGO6: Reflex Level accommodates instrument operations that do not require advanced processing such as adjustments to maintain proper speed, altitude, and heading as well as quick interpretation of data. Control theory is used to ensure control stability and performance optimality.

EGO7: Integrity and Health State Management concerns itself with diagnostics, cybersecurity, emergency instrument availability, and redundant sensors/instruments.

Ground-Truth Ego Environment

The *ego(ENV)* analysis once again is narrowed to be within the physical scope of the operating airspace. Thus, it contains the same entities as before, except for the pilots replacing the autopilot.

ENV3: Environment Plans consequently contain the pilots whose objective is the climb to altitude 33,000 ft MSL, in addition to the ATC communicator and the other aircraft.

ENV4: Environment Strategies hold the same plans of ATC and other aircraft as before as well as the pilot's plans to call ATC for clearance, oversee system operations, and input data into technical systems that might be congruent or conflicting to autopilot expectations.

ENV5: Environment Capabilities and Dynamics now consist of the pilot's performance, which is based on experience and physical/cognitive ability. Furthermore, the autopilot has a different relation to the ATC and other aircraft, because it relies on their technical systems to receive and correctly interpret its transmitted data. Using only a single frequency for communication does not allow simultaneous communication, which could result in a riskier environment.

ENV6: Environment Classification, State, and Belief for the autopilot are determined through the geospatial orientation parameters: flight management system, pilot inputs and controls, accessory instruments, the airspace, weather conditions, and altitude.

ENV7: Uncontrollable EGO State, Capabilities, and Dynamics include instrument malfunction, erroneous sensors, incorrect inputs, and rare/unexpected events.

3 THE INTERACTION METAMODEL

We will now outline general principles for the information exchange between actors in a human cyber-physical system, focusing on the interaction of humans with cyber-physical systems.

The presentation of RA(HCPS) in the companion paper [14] abstracts completely from the realization of the interaction between different systems. It suggests what categories of information are exchanged between actors but leaves open how this is achieved. In this section, we sketch requirements and concepts on and of a reference architecture for multi-party man-machine interaction within safety-critical real-time systems situated on some level of the aggregation hierarchy. We assume, much as in the case study in Section 2, that humans involved in the interaction are professionals or semi-professionals. They thus have undergone training to understand normative and regulatory context as well as the capabilities and limits of the technical systems with which they are interacting and have the necessary capabilities to interact. We strive for identifying general principles underlying such interactions, which are generic and go beyond individual domain-specific applications, and which are of a sufficiently fundamental nature to justify their integration into a reference architecture. The focus is thus on functional and extra-functional requirements for interaction models, which can guide implementation choices to design the human-machine technology interaction. We focus on human CPS interaction (rather than interaction between multiple CPSs) and ask what type of information exchange must be supported on what level so that humans and systems can cooperate as a team, what is the criticality of exchanged information, what are timing requirements for such interactions, and how we can communicate highly critical information in a limited time window—in spite of the many sources of distorted perception addressed in [14] and revisited below. We motivate and outline challenges for the interaction metamodel, highlight shared stumbling blocks, and illustrate shared design principles, which fundamentally rest on established ontologies for particular application classes. We anticipate a key role of virtual twins of both human and technical cooperation partners as well as of the controlled system and its environment for designing a suitable communication.

3.1 Meta Requirements

Throughout this article, we assume that the information exchange involves members of some team, i.e., a set of systems (humans and technical systems), that have agreed to cooperate for some given time period to achieve a certain goal state. Section 2.1 of [14] defines this as "*groups of systems* operating in tight interaction due to (relatively) close physical proximity." Teams are created dynamically for an agreed-upon period of time either by systems in a physical neighborhood on the same level in the aggregation hierarchy (such as a team of medical equipment, nurses, and doctors in an intensive care unit, or the team formed by the driver with her car) or by becoming a member of a super-system at a higher aggregation level—in which case the system is expected to both comply to and benefit from the rules and services of this super-system (such as a patient entering a hospital or a car entering a highway control system). Each negotiation leading to a formation of a team defines which layers of the participating systems will be involved in the cooperation; e.g., in cooperative driving, cars may agree to cooperate on the maneuver level by forming a coalition for the duration of the time it takes to pass an object blocking the road ahead, by subscribing to the rules of a traffic flow coordination system to cooperate on the strategy level for safe minimal energy consumption crossing of the intersection ahead, or by cooperating on the planning level to form a platoon of trucks minimizing fuel consumption by sharing and optimizing plans to reach the destinations of the individual trucks forming the platoon.

We also assume that all systems are embedded into a super-structure (by default the nation in which the system is deployed), where they share ethical, societal, and psychological principles as well as regulatory, legal, and physical constraints—thus all systems agree on the moral-system layer and the reflection layer. Technical systems crossing borders or operating across borders, such as an automated truck traveling from the traffic system of one country to another, have to be designed in a way allowing customization to such principles and rules and necessitate the creation of internationally agreed-upon ethical and legal standards. Real-time requirements of actions on the reflex level are such that team formation at the reflex layer is not possible—any system will react on its own to environmental threats at the reflex level within its specific temporal capabilities (delay times). However, since all systems share common ethical and regulatory principles, reflexive or instantaneous actions shall typically reduce risk for systems that are not part of their own team.

We start out by identifying, per layer of the ego systems (see the companion paper [14]), the classes of information exchanged between systems of a team and two non-functional requirements that we call criticality and time stability (see Figure 1).

We define the criticality of the exchanged information as a measure of the level of risk to not achieve the layer-specific goals of the team. To determine such criticality, methods for analyzing the maximal risks in violating team goals when this information is not perceived at the receiver side can be used. A further metric is the rate with which states on the next higher levels have to be changed to ensure system stability. Time stability defines the maximal period before which an updated value must be transmitted, a term also referred to in the real-time systems community as the periodicity of messages. A third non-functional requirement, not shown in Figure 1, is the *deadline* for the end-to-end latency between emission and perception of the message. A key challenge in HMI design is to ensure that critical information is registered, processed, and correctly acted upon within the given deadlines/time limits (see also following section). Additional non-functional requirements are related to criticality—whenever descriptive beliefs are exchanged between systems, it is indispensable that the sender attributes such beliefs with his or her own estimated level of confidence and precision of the exchanged belief, and that the quality of these estimates matches the criticality level of the exchanged belief.

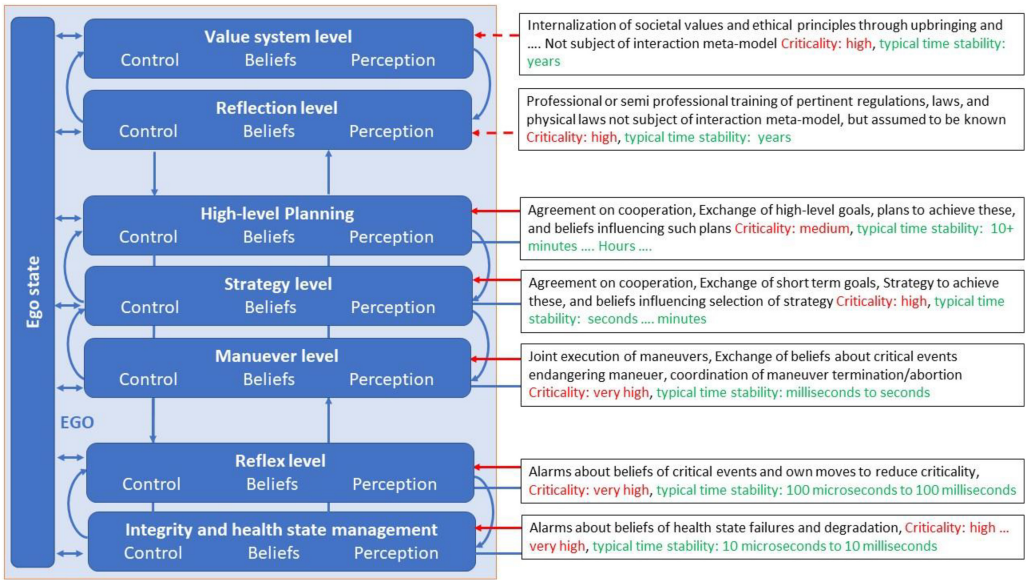

Fig. 1. Type, criticality, and time stability of information exchange at different levels of ego metamodel.

Many obstacles can prevent humans and technical systems from working as a team. Klein et al. [22] have identified the following "Ten Challenges for Making Automation a 'Team Player' in Joint Human-agent Activity," which must be addressed in any implementation and constitute requirements on the expressiveness of our reference architecture. In the terminology introduced in Section 2.1 of Part I [14], we referred to *teams* as "*groups of systems* operating in tight interaction due to (relatively) close physical proximity (e.g., an operator in a control room interacting with the CPS controlling traffic flow, a platoon of trucks, a team of doctors and nurses in an emergency room supported by a multitude of medical devices)."

Challenge 1:
To be a team player, an intelligent agent must fulfill the requirements of mutual predictability, directability, and common ground to engage in common-grounding activities.

Challenge 2:
To be an effective team player, intelligent agents must be able to adequately model the other participants' intentions and actions vis-à-vis the joint activity's state and evolution. For example, are they having trouble? Are they on a standard path proceeding smoothly? What impasses have arisen? How have others adapted to disruptions to the plan?

Challenge 3:
Human-agent team members must be mutually predictable.

Challenge 4:
Agents must be directable.

Challenge 5:
Agents must be able to make pertinent aspects of their status and intentions obvious to their teammates.

Challenge 6:
Agents must be able to observe and interpret pertinent signals of status and intention of other agents.

Challenge 7:
Agents must be able to engage in goal negotiation.
Challenge 8:
Support technologies for planning and autonomy must enable a collaborative approach.
Challenge 9:
Agents must be able to participate in managing attention.
Challenge 10:
All team members must help control the costs of coordinated activity.

We note that this type of tight cooperation in a team differs fundamentally from the much looser control of humans (and CPS) acting as members of a super-structure, who all share the overall objectives of this super-structure; e.g., drivers will adapt their overall driving style to mostly comply to existing traffic regulations, but repeatedly violate these for short-term personal reasons. In super-structures with more rigidly enforced shared objectives, this type of randomness in behaviors can actually drastically reduce the demand for explicit coordination and cooperation. Consider the emergency rescue scenario of the companion paper [14]: in such an emergency situation, the deployed emergency rescue staff will not have received detailed instructions as to what specific role and resulting tasks to take over. Instead, the mere fact that such rescue staff will arrive at different points in time, the training they have experienced, and the objectives they share will cause them to fill the most needed roles and tasks on the spot, as they arrive. Thus, the randomness associated with the arrival processes, together with shared objectives and professional training, replaces the costly amount of time needed to negotiate the distribution of roles and tasks first. This scenario thus combines both elements of predictability and randomness in a synergetic way: it shows predictability in, that every member of the rescue team will pick the most urgent role and task upon arrival, but exactly what role and what task is not predetermined but is subject to the randomness of the arrival process.

Communication and cooperation can be hindered by distorted world views as listed below. Note that some of these points are valid for both humans and intelligent technical agents. For technical systems, even for directly observable real-world entities, the representation of the world will differ from reality due to effects ranging from noise at the physical level to inherent tradeoffs in excluding false negatives versus guaranteeing high detection rates for object extraction out of video stream images.

- For humans, perception can be distorted because of lack of attention, limited processing capabilities, or the influence of human states such as stress and fatigue.
- Misguided pre-filtering of sensory information: The necessary pre-filtering of sensory information is influenced by the state of the ego system and its currently pursued tasks and may miss relevant information.
- Inadequate world models: Even with perfect sensory systems, actions of the system will fail drastically if perceptions are interpreted in an inadequate world model.
- Limited perception bandwidth: Perception is inherently limited by the perception bandwidth.
- Failures: These effects are aggregated by failures in communication and sensory systems, which, without proper error detection and recovery mechanisms, can lead to arbitrary large gaps between perceptions of reality and reality itself.
- Unawareness of health state of perception system: The health state of a system is uncontrollable, hence—as part of env(ego)—not directly observable. Thus, incorrect beliefs about perception components may lead to distorted perception.

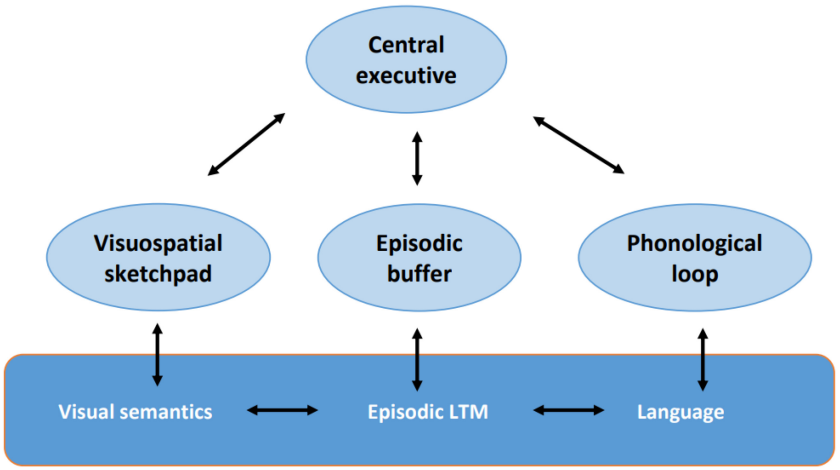

Fig. 2. The working memory model by Baddeley and Hitch [3], taken from Banales et al. [5].

3.2 The Fundamental Essence

3.2.1 The Role of Descriptive Beliefs in Human Ego Systems. In the above considerations, several concepts from cognitive psychology (e.g., attention, memory, schemas) have been used. Here we outline a basic and general model of human information processing that builds on well-accepted models in cognitive psychology [6]. This is also relevant for the concept of descriptive beliefs introduced in the companion paper (c.f. Annex 1, [14]), relating the EGO system's internal states with ground-truth ENV states. In humans, descriptive beliefs can guide what aspects of the self or the environment are attended to and perceived and which information is processed and stored in memory. Long-term memory stores the acquired (learned) schemata of situations and actions, which are used to guide attention, information processing, and memory formation [28]. It can be assumed that the main part of cognitive processing of information for a later response selection transpires within working memory.

A highly influential model of human information processing in working memory, introduced by Baddeley and Hitch [3], proposes that information and schemata currently processed are held and manipulated in one of three modality-specific temporal stores (see Figure 2): a visuo-spatial sketchpad for visual input, a phonological loop for auditory input, or an episodic buffer for information about events or experiences. Information processing in working memory can be supported and structured by schemata retrieved from long-term memory. The existence of a short-term sensory buffer with nearly unlimited capacity is assumed but not considered as part of working memory. Each of the working memory stores is greatly limited in its capacity (7 ± 2 information chunks [25]), and schemata can help to increase the efficiency of information processing [28]. It can therefore be assumed that only a small minority of memory contents and descriptive beliefs are represented in working memory at each moment to support the interpretation and evaluation of sensory information.

The allocation of information to memory stores, as well as monitoring and coordination of the entire system, is handled by a central executive. The central executive can be conceived as the manager of working memory, which is also connected to long-term memory from which it can retrieve information about task coordination. The central executive holds and manages goals, decides which information requires attention, and resolves conflicts. The central executive is also responsible for selecting information to be stored in long-term memory for the formation or updating of

new and old descriptive beliefs. Different categorizations of information stored within long-term memory have been proposed, e.g., the difference between declarative or procedural knowledge [2]. For technical systems, Crowder and Friess argue [12]:

"In order for an Artificially Intelligent System (AIS) to be truly autonomous, we must provide the system with the abilities to acquire, categorize, classify, store, and retrieve information and knowledge, and provide abilities to infer or reason about the knowledge that it has stored. This drives the need for memory types that are similar to the different memories in the human brain (Crowder and Friess, 2010, 2011 [10, 11]):

- Sensory Memory (where raw, unprocessed information from sensors is buffered and initial pre-processing is accomplished)
- Short-Term Memory (called 'working memory,' where new information from sensory memory is stored while it is processed and 'reasoned about')
- Long-Term Memory (where permanent knowledge is stored through rehearsal, encoding, and memory association occurs)
- Emotional Memory (memories about experiences, events, or information that cause 'stress' within the AIS)"

We think that these theories illustrate our view that descriptive beliefs, reasoning, and memory in technical systems show many similarities to the theories about the mechanisms of information processing in humans. In our metamodel, we assume that emotional memory is associated in the lowest reflex layer of the ego system and that descriptive beliefs at higher layers are switched between memory types. Any conscious act of the ego system (in all higher layers) requires the relevant part of its current descriptive belief to be in working memory.

3.2.2 Ontologies. Exchanging information entails that the team players "use the same language." To this end, the interaction metamodel requires for each type of team and for each layer of cooperation the existence of an ontology that defines completely the type of information exchanged between team players at the chosen layer of cooperation. This includes relevant environmental artifacts, their static and dynamic attributes, and relations between artifacts. Static and dynamic attributes of entities in an ontology can be defined using concepts such as class definitions, which can define dynamic capabilities of the entity (such as maximal speed and/or acceleration) as well as statistical models of typical behaviors in well-defined environments (e.g., the typical movements of pedestrians when wanting to cross a street). Ontologies will typically be partially ordered, where the ordering relation corresponds to specialization. Ontologies of neighboring layers are interrelated by *aggregation* (resp. *explication*) relations—a key requirement to support "zooming in." *Explication* supports unfolding the representation of a critical system state on a lower level (such as zooming in to a highly congested intersection to identify that it was a failure of a traffic light that caused the congestion). *Aggregation* allows to abstract from detailed information so as to be able to easily identify systems requiring immediate attention. The **ecological interface design (EID)** (see below) will try to identify the most abstract layer at which information is to be communicated, along the hierarchy shown in Figure 3.

We distinguish between the elements of such ontologies and their representation in information exchange. Both deadline and bandwidth requirements typically require the representation to be non-textual, and thus they require the introduction of visual, haptic, or acoustic idioms or metaphors for representing the artifacts of the shared ontology. Specifically, the metamodel requires the specification of a representation view of each artifact, attribute, or relation in the ontology.

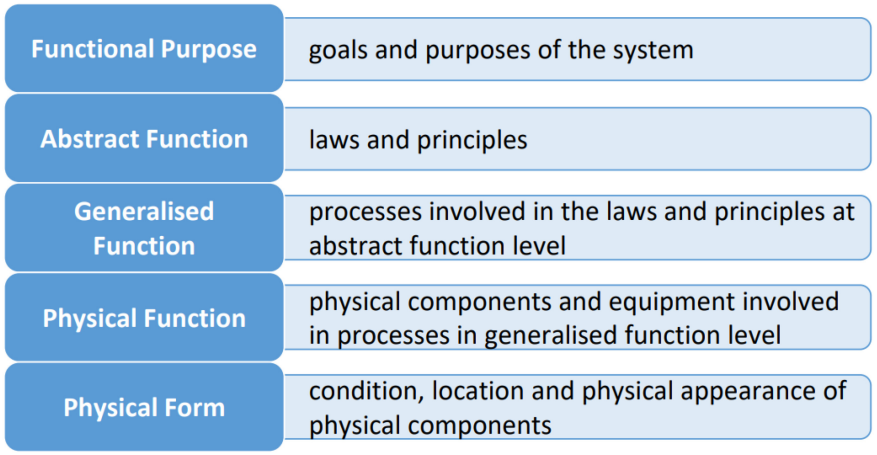

Fig. 3. Abstraction hierarchy of EID framework.

EID is a theoretical framework developed by Vicente and Rasmussen for designing interfaces for complex human-machine systems [31, 8]; see Figure 3. "Unlike classical part-whole hierarchies, the levels in the functional abstraction hierarchy are connected via means-end links: The functional purpose is realized by making use of the physical laws (abstract function); therefore, these laws are realized by applying certain generalized functions in specific sequences; the generalized functions are realized by physical system components, which are described with regard to their physical functions and their physical forms creating the physical function. This structure shall support operators in problem-solving. If the system is working correctly, the principles and general conditions on each level, as well as their relationships, are fulfilled; in case of failure, some of them are violated. The abstraction hierarchy allows to systematically "zoom" into the system via the means-end links. Often a failure becomes apparent at the system surface by the fact that a particular purpose of the system is not fully obtained any more. Starting from there it is possible to detect the physical laws that are violated due to the fact that particular generalized functions are not fulfilled, which can be further tracked down to the loss physical functions and thus to malfunctions of one or more system components. The "zooming" from abstract levels to more concrete levels allow to focus on those physical components, which are involved in producing the violated functions. In this way, the abstraction hierarchy allows goal-oriented and efficient problem-solving, and thus, countermeasures can be initiated faster" (cited from [18], p. 27).

We also assume for each super-structure the existence of a background ontology that is shared between all systems belonging to this super-structure. Such background ontologies are typically defined on the level of homogenous collections of cyber-physical systems, as defined in Section 2.1 of the companion paper [14], such as rules guiding traffic or regulations regarding emissions. Highly regulated domains such as aerospace rely on international standards defining such ontologies. Initiatives such as A.U.T.O. pushed by the German Lighthouse Project on Validation and Verification for highly autonomous driving[1] are striving to define such an ontology accepted throughout the German automotive industry.

3.2.3 Representation Principles for Ontologies. Promising approaches to represent complex information under real-time constraints include the use of augmented reality and virtual reality. Before discussing these principles, let us use an example taken from Sonntag and Möller [29] of

[1] https://www.vvm-projekt.de/en/

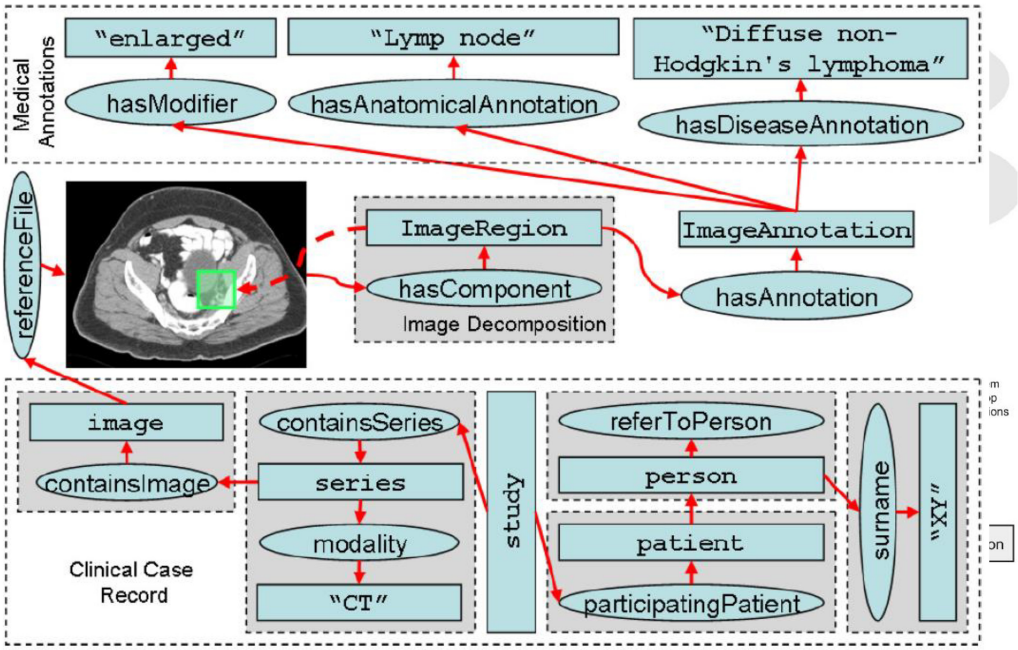

Fig. 4. Domain-specific ontology for analyzing medical images.

the medical domain to highlight how applying domain-specific knowledge makes it possible to create easy-to-understand interfaces that minimize risks of misconceptions and misunderstandings. A combined multimodal user interface for the semantic annotation and retrieval of medical images is depicted in Figure 5 and based on a domain-specific ontology shown in Figure 4.

For such cases of visual representations of artifacts to be communicated, Figure 6 reminds us of the complex processes that humans must be able to perform in real time. The processes range from raw perception to understanding on the cognitive level, planning and decision making, and induced follow-up interactions.

In Figure 6, the User itself should be seen refined as shown in Figure 7.

An important benefit is the human ability to reason about the data and extract higher-level knowledge, or insight, beyond simple data transfer [9]. This benefit enables users to infer mental models of the real phenomena represented by the data [18]: "Insight defines what the human operator should intuitively apprehend when looking at the visual form—e.g. quantitative value of an information, certainty of an information, or to which category or mode the information belongs[2]." Case studies such as the one discussed in the previous section have shown that additional confirmatory information is required. This insight is often needed by human operators when dealing with sensor values. In this case, the certainty of the correctness of the values is often needed as insight by the human operator[3]." This insight is not identified by Denzau et al. [15].

In general, Denzau proposes the following requirements on the visualization pipeline:

This mapping should be computationally done by some function F by taking the dataset as input and generating the visual form as output [15]. This function F must meet four important requirements:

[2] [18], p. 50.
[3] [18], p. 54.

- Computable F can be computed by some algorithm.
- Invertible[4] F must be invertible by some function to ensure that the visualization is not ambiguous or not understandable by the user looking at the visualization.
- Communicable F must be known by the user so that he or she can understand the visualization and interpret the shown information in the right way. Based on Denzau et al. [15], this is a learnability issue.
- Cognizable F should be easily perceivable by the user and should minimize his or her cognitive load for decoding and thereby interpreting the visualization[5].

In addition to explicit communication, implicit communication also plays a role. Machine behavior is easier to understand and predict (see Team Player Challenge 3 from above) if it is based on consistent criteria and principles that are applied by humans too. Furthermore, the speed of machine operation must be adapted to the processing speed of humans in order to be understandable. Consequently, in order to be a good "team player," machines have to trade off between maximal efficiency and optimal predictability for the sake of the safety of the overall human-machine system [4].

The process of information processing as described in Figure 3 is understood as bottom-up processing from incoming information units to representation of complex meaningful concepts. In parallel, a top-down process based on experience, preexisting knowledge, and goal-directed intentions influences attentional resources necessary for the bottom-up processing. Visual behavior, fixation points, and visual search patterns can be used to understand the current top-down processes better. Certainly, top-down processes enable the application of anticipative strategic behavior and the focused search for necessary information. It can also lead to missed information or biased interpretation.

For a team situation, this fact plays a dominant role. It is of central importance to align the strategic top-down processes of the team members either to instantiate an obvious and complementary situation for all members or to avoid contradicting decisions due to different levels of experience and/or intentions.

3.2.4 Highlight Exactly Those Aspects of the Systems That Are Pertinent to the Current Decision Process. We suggest to consider the layers of the ego systems metamodel from the perspective of the EID abstraction hierarchy (cf. Figure 3). Within this abstraction hierarchy, the layers are linked to each other by a goal-means relation: the functional purpose is implemented by means of the laws and principles of the abstract function, which are implemented by processes of the generalized function; these functions are realized by physical components, which are defined by their physical functions and on the lowest level by their physical attributes contributing to the functions.

Behavior and phenomena on higher levels can be understood by looking at the contributing means and effects on lower levels. The abstraction hierarchy supports zooming in to a complex system. For example, an error often surfaces on the highest level through a functional purpose that is not performed correctly. To understand the reason, violated abstract functions on the next lower level and finally physical forms can be analyzed. Zooming in allows us to focus on those parts of the complex system that are relevant for the behavior or phenomenon in question.

EID provides guidelines for deriving the form of information presentation for the relationships within the abstraction hierarchy. EID refers to the three levels of information processing as defined

[4]We note that such invertibility does in no way imply that the function F is injective. Instead, the capability to recover the key situational dependent aspect of the original raw data must rest on further context data and training.
[5]Slightly adapted from [18], p. 44.

by Rasmussen [27]: skill-, rule-, and knowledge-based information processing. Rassmussen defined these levels based on the involved information types, knowledge structures, and mental resources:

- Skill-based information processing is activated in cases where information is perceived as "signals" within a space and time continuum. These signals activate continuous motoric regulating behavior that is performed unconsciously without much thinking involved, e.g., steering a car within the delimiters of a street.
- Rule-based information processing is activated in cases where information is perceived as "signs." The signs activate rules that are performed partially conscious, because a choice between alternative rules has to be made by actively looking for and considering further information.
- Knowledge-based information processing is activated in cases where information is perceived as "symbols." The situation is new and no learned behavior or process is mentally available. The situation has to be analyzed consciously, and solutions have to be found, e.g., by mental simulation or analogy.

Skill-based information processing requires the fewest mental resources and knowledge-based processing the most mental resources. The taxonomy indicates that the level that is activated in a concrete situation strongly depends on the form of information processing: signals, signs, symbols. Human-machine interaction should support processing on the lowest level, which is possible considering the task as well as the experience of the human.

The following guidelines can be derived for the design of the human-machine interaction (see [31]):

- Supporting skill-based processing: The human should be given the ability to manipulate objects on a display similar to the interaction with real objects. In order to activate regulating motoric behavior, the spatial and temporal attributes have to be directly relatable to the corresponding real objects. The display should directly show how the manipulation of displayed physical objects changes the performance with regard to the functional purpose and how this influence is caused by the underlying laws, principles, and functions on all layers of the abstraction hierarchy. Principles from geometry can be applied in order to show all these dependencies using graphical signals (cf. [17]). If possible, a temporal highly synchronized multimodal information presentation is preferred to just visual information presentation.
- Supporting rule-based processing: In order to make correct choices, the display has to show the signs that are relevant for the applicable rules. Defining rules and showing signs of the abstract and generalized function allow a deeper understanding of system (cf., e.g., [1]). Contrarily, monitoring and control tasks are often defined and supported on physical levels. In many cases this does not reveal how the rules and resulting actions of humans contribute to overall purposes of the system. Rules should be defined on all levels and the links between them should be visually highlighted on corresponding displays.
- Supporting knowledge-based processing: The display should support deep analysis of the system by presenting the full abstraction hierarchy (see [7]) and by allowing easy zoom-in and zoom-out of the dynamic properties of the system. This allows the human to explore the system and anticipate the effects of actions on all levels as required in order to find a solution.

These guidelines provide further requirements for the information content (in addition to those in Figure 5) and for the structure of the information presentation. Finally, the metamodel also suggests a way to derive concrete presentations of information. We focus these guidelines to visual

3. U: "Show me the internal organs: lungs, liver, then spleen."
4. S: Shows patient images according to referral record.
5. U: "Annotate this with lymph node enhancement" (+ pointing gesture on region); "so *lymphoblastic*" (expert finding, additional disease annotation (ICD-10)).
6. S: "Region has been annotated."
7. U: "And replace the characteristic of the other by RadLex: shrunken."
8. S: "Region characteristic has been updated." → The radiologist switches to another patient (for illustration purposes with a broken finger) and asks for a summary in this additional retrieval stage.
9. U: "Give me a summary of this patient." (retrieval stage)
10. S: "This is a summary of the fracture: . . . "
11. S: "Five corresponding CTs will be displayed." → The radiologist can now switch again to the differential diagnosis of the suspicious case together with a second medical expert (for the first patient), where the case is examined again and the image annotations can be completed.

Fig. 5. A combined multimodal user interface for analyzing medical images.

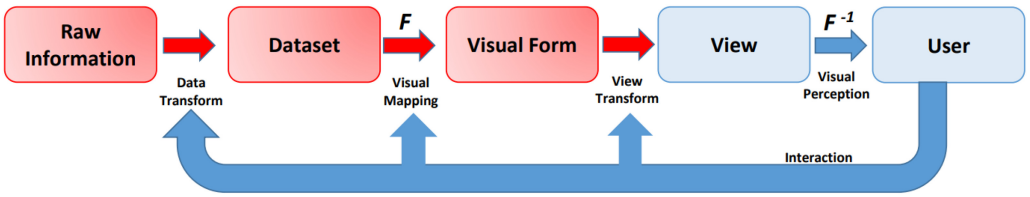

Fig. 6. Visualization pipeline (based on [15] and on adaption by [9]).

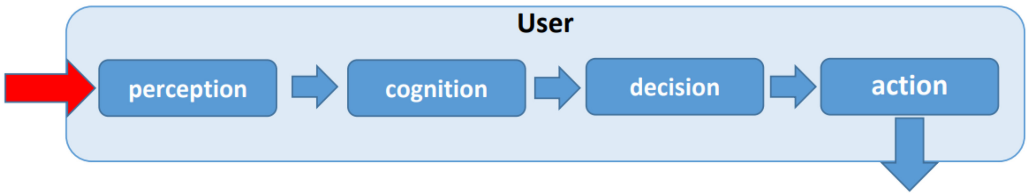

Fig. 7. Refinement of User model.

presentations and refer to the KONECT method of Ostendorp et al. [26]. KONECT is a comprehensive approach to deriving visual representations of information based on an analysis of the purpose of the communicated information: for which task does the human need the information? Harre defines "insight" as follows: "what the human operator should intuitively apprehend when looking at the visual form—e.g., quantitative value of an information, certainty of an information, or to which category or mode the information belongs." KONECT supports choosing adequate visual attributes for various insights and for combining them in one combined form, a visual glyph. KONECT is deeply rooted in psychological theories like Gestalt theory [23].

Time constraints are another factor that play an important role for the selection of the communication strategy under severe time constraints, and in complex environments, it might be advisable

to support skilled-based processing, i.e., only presenting alternative choices of actions that are available. Note, however, that this requires a deep and reliable understanding of the situation and the admissible paths of action by the CPS.

If time constraints exist but the CPS lacks sufficient understanding of the situation to plan admissible paths of action, it might be more favorable to support rule- or knowledge-based processing. However, under time constraints, it is important to factor the severe information processing bottleneck in human cognition into the implementation of the communication strategies. In this case, the CPS's communication strategy should reduce the information to the human to build a plan for action to the mere necessary or at least highlight the important artifacts/actions in the environment. The selection of this information can be based on a virtual twin, i.e., simulation of the environment based on redundant sensor data that is run in parallel to the CPS and serves as its basis for strategy and action planning. Such a representation of the environment can serve as a basis for communication of very focused information to a trained human (we assume trained humans in this article) even when the CPS is not capable of providing a safe plan of action (see, e.g., Waymo car killing bicyclist due to rapidly changing object classifications). Since the creation of sufficiently faithful virtual twins is seen as a central building block for establishing safety for highly autonomous systems as well as over-the-air upgrades of key functions, it is generally expected that such virtual twins will be available within the next few years (cf. Presentation by Siemens Mobility, SafeTRANS Industrial Day, Nov. 28, 2022). Such virtual twins will be an ideal basis for highlighting exactly those aspects of the systems that are pertinent to the current decision process.

3.3 Can We Validate That the Relevant Information Has Actually Been Understood by the Human Operator?

Probing cognitive or physiological indicators for misunderstanding has several problems and cannot guarantee that the world representation of the human and the cyber-physical system is aligned. Misalignment may simply have gone unnoticed. In humans, understanding of information can be probed actively or passively.

Probing predictors of limited cognitive abilities (like mental workload, emotional arousal, and fatigue) are probabilistic in the sense that as noise degrades empirical measurements, they can predict a state only up to a level of uncertainty. Moreover, these states are only modulators of the cognitive abilities and not direct indicators of understanding of a message with respect to content and consequences.

Active probing of information understanding in humans can be implemented by asking them to acknowledge or copy messages. This was demonstrated in the case study in Section 2. As an example, the captain asked twice when he was not sure he correctly understood the new radio frequency and copied it back after it was repeated by the first officer.

Overt behavior like eye-scanning patterns, eye blinks, facial expressions [21]—together with sub-lexical information in verbal expressions and physiological measures like heart rate variability—can also provide information about the cognitive and emotional states or fatigue of human operators. These probes are inherently passive as no active communication of the inferred state is necessary. Instead, it is deduced from observational data. In principle, these operator states will also be available at the brain level. Brain decoding studies in the field of neuro-ergonomics [24] indicate that some cognitive and emotional states are recognizable in specific brain activation patterns that are accessible with non-invasive measurement techniques. For example, Unni et al. [30] showed that the current level of the working memory load of drivers can be decoded from brain activation. Moreover, Ihme et al. [21] demonstrated that increased levels of frustration, indicating loss of control and anticipated failure to reach a goal, can be decoded from brain activation measurements. Moreover, decoding from brain activation is more accurate than decoding from

facial expressions as the latter are more variable. As a proof of concept, Damm et al. [13] showed that the safety of a human cyber-physical system interaction can be improved by adapting the behavior of the cyber-physical system to the increased frustration level of the human, even when the information about the human state is only probabilistic. A weakness of the passive measures is that they are often only indirect indicators of states and may be influenced by several demands or internal states, although they clearly differ with respect to their specificity, with brain activation patterns being likely the most specific but at the same time the hardest to obtain. However, in order to obtain insights in the exact content of internal information representation, active probing appears to be necessary because none of the passive techniques allow direct inferences about the functional processes of information processing in relatively unconstrained realistic settings.

Digital twins offer a promising approach to open up the box and to provide insights into the information processing in the human cyber-physical system. This requires that the human model adequately represents the human in the most relevant cognitive operations, as well as sensory and motor capabilities, at least in the limits of the task requirements. Such human models have long been used for the evaluation of human user interfaces in human cyber-physical system interaction. Recently, Held et al. [19] have shown that such a model can adequately reproduce behavioral consequences of cognitive task interference during driving, and at the same time, it can provide time-resolved insights into how symbols are processed and conflicting goals functionally can interfere at the level of cognitive processing. By their nature, the insights provided by digital twins are probabilistic and only correct to the extent that the model is adequate. However, by comparing the behavior of humans and digital twins when they perform the same tasks, some indicators for the adequacy of the model can be derived. We conjecture that digital twins performing the tasks as the human operators are a promising route to integrate hypotheses about human states into the interaction of cyber-physical systems with human operators.

4 CONCLUSION

In this article, we propose an interaction metamodel of human cyber-physical systems, with focus on the interaction of humans with cyber-physical systems. The increasing automation in such systems highlights the need to design them in a way that maximizes the probability that human operators perceive key artifacts and comprehend their roles in the system to ensure that mutual beliefs about goals, plans, strategies, and perceptions of the environment are sufficiently consistent and sufficiently precise. With an example from aviation, we demonstrate the usefulness of the reference architecture for the analysis of the communication between agents to identify potential communication problems. We use the proposed layered metamodel to define the criticality of the exchanged information as a measure of the level of risk to not achieve the layer-specific goals of the team, and we discuss obstacles that can prevent humans and technical systems from working as a team. We suggest a shared layered ontology that is parallel with the structure of the metamodel as a key element for mutual understanding and error checking in a human cyber-physical system. These abstract analyses need to be complemented by design principles that consider the expectations and limited information processing capacity of humans. We propose that human-user interfaces should account for processing speed and types of information conveyed, and the interfaces should support goals-ends insights. For the future, we anticipate a role for digital twins of both human and technical cooperation partners to overcome the partial opacity of their internal states and to ensure suitable communication.

REFERENCES

[1] Matthijs H. J. Amelink, Max Mulder, Marinus M. Van Paassen, and John Flach. 2005. Theoretical foundations for a total energy-based perspective flight-path display. *International Journal of Aviation Psychology* 15, 3 (2005), 205–231.

[2] James R. Anderson. 1976. *Language, Memory, and Thought.* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

[3] Alan D. Baddeley and Graham J. Hitch. 1974. Working memory. In *The Psychology of Learning and Motivation: Advances in Research and Theory*, G. H. Bower (Ed.). Vol. 8. Academic Press, New York, 47–89.
[4] Lisanne Bainbridge. 1983. Ironies of automation. *Automatica* 19, 6 (1983), 775–779.
[5] Erin Banales, Saskia Kohnen, and Genevieve McArthur. 2015. Can verbal working memory training improve reading? *Cognitive Neuropsychology* 32, 3–4 (2015), 104–132.
[6] Guy A. Boy. 1998. *Cognitive Function Analysis.* Ablex, Distributed by Greenwood Publishing Group, Westport, CT.
[7] Caterine M. Burns. 2000. Putting it all together: Improving display integration in ecological displays. *Human Factors* 42 (2000), 226–241.
[8] Caterine M. Burns and John Hajdukiewicz. 2004. *Ecological Interface Design.* CRC Press, Boca Raton, FL.
[9] Stuart K. Card, Thomas P. Moran, and Allen Newell. 1983. *The Psychology of Human–computer Interaction.* Lawrence Erlbaum Associates, Hillsdale, NJ.
[10] James A. Crowder. 2015. *Ontology-based Knowledge Management.* Society for Design and Process Science.
[11] James A. Crowder and Shelli Fries. 2011. Metacognition and metamemory concepts for AI systems. https://www.researchgate.net/publication/235219069
[12] James A. Crowder and Shelli Friess. 2012. Artificial neural sensory/short-term/long-term/emotional memory integration for autonomous AI systems. In *AIAA SPACE 2012 Conference & Exposition.*
[13] Werner Damm, Martin Fränzle, Andreas Lüdtke, Jochem W. Rieger, Alexander Trende, and Anirudh Unni. 2019. Integrating neurophysiological sensors and driver models for safe and performant automated vehicle control in mixed traffic. *2019 IEEE Intelligent Vehicles Symposium*, IV 2019, Paris, 82–89.
[14] Werner Damm, David Hess, Mark Schweda, Janos Sztipanovits, Klaus Bengler, Bianca Biebl, Martin Fränzle, Willem Hagemann, Moritz Held, Klas Ihme, Severin Kacianka, Alyssa J. Kerscher, Sebastian Lehnhoff, Andreas Luedtke, Alexander Pretschner, Astrid Rakow, Jochem Rieger, Daniel Sonntag, Maike Schwammberger, Benedikt Austel, Anirudh Unni, and Eric Veith. 2023. A reference architecture of human cyber-physical systems .part I: Fundamental concepts. *ACM Trans. Cyber-Phys. Syst.* (October 23 2023), 32 pages. https://doi.org/10.1145/3622879
[15] Arthur T. Denzau, Douglass Cecil North, and Ravi K. Roy. 2006. Shared mental models. In *Neoliberalism.* Routledge, 14.
[16] Lida David and Jan Maarten Schraagen. 2018. Analyzing communication dynamics at the transaction level: The case of Air France Flight 447. *Cognition, Technology & Work* 20 (2018), 637–649. https://doi.org/10.1007/s10111-018-0506-y.
[17] John M. Flach and Kim J. Vicente. 1989. Complexity, difficulty, direct manipulation and direct perception. Technical Report EPRL-89-03. Engineering Psychology Research Lab, University of Illinois, Urbana-Champaign, IL.
[18] Marie-Christin Harre. 2019. *Supporting Supervisory Control of Safety-critical Systems with Rationally Justified Information Visualizations.* Ph.D. Thesis, Carl von Ossietzky Universitaet Oldenburg, Germany.
[19] Mortiz Held, Jochem W. Rieger, and Jelmer P. Borst. 2022. Multitasking while driving: Central bottleneck or problem state interference? *Human Factors* (2022), https://doi.org/10.1177/00187208221143857
[20] Edwin Hutchins. 1995. Distributed cognition in an airline cockpit. *Cognitive Science* 19 (1995), 265–288.
[21] Klas Ihme, Anirudh Unni, Meng Zhang, Jochem W. Rieger, and Meike Jipp. 2018. Recognizing frustration of drivers from face video recordings and brain activation measurements with functional near-infrared spectroscopy. *Frontiers in Human Neuroscience* 12 (2018), 327.
[22] Gary Klein, David D. Woods, Jeffrey M. Bradshaw, Robert R. Hoffman, and Paul J. Feltovich. 2004. Ten challenges for making automation a "team player" in joint human-agent activity. *Intelligent Systems*, IEEE, Vol. 19, 91–95. 10.1109/MIS.2004.74
[23] Riccardo Mazza. 2009. *Introduction to Information Visualization.* Springer, London.
[24] Ranjana K. Mehta and Raja Parasuraman. 2013. Neuroergonomics: A review of applications to physical and cognitive work. *Frontiers in Human Neuroscience* 7 (2013), 889.
[25] George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (1956), 81–97.
[26] Marie-Christin Ostendorp, Thomas Friedrichs, and Andreas Lüdtke. 2016. Supporting supervisory control of safety-critical systems with psychologically wellfounded information visualizations. In *Proceedings of the 9th Nordic Conference on Human-computer Interaction (NordiCHI '16).* ACM, New York, NY, 11:1–11:10.
[27] Jens Rasmussen. 1983. Skills, rules, knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics* 13, 3 (1983), 257–266.
[28] David E. Rumelhart. 1984. Schemata and the cognitive system. In *Handbook of Social Cognition.* R. S. Wyer Jr. and T. K. Srull (Eds.). Lawrence Erlbaum Associates Publishers, 161–188.
[29] Daniel Sonntag and Manuel Möller. 2010. A multimodal dialogue mashup for medical image semantics. *IUI'10: Proceedings of the 15th International Conference on Intelligent User Interfaces*, 381–384. https://doi.org/10.1145/1719970.1720036

[30] Anirudh Unni, Klas Ihme, Meike Jipp, and Jochem W. Rieger. 2017. Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: A realistic driving simulator study. *Frontiers in Human Neuroscience* 11 (2017), 167.
[31] Kim J. Vicente and Jens Rasmussen. 1992. Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man and Cybernetics* 22 (1992), 589–606.

Received 27 January 2023; revised 7 July 2023; accepted 2 August 2023