A REFERENCE ARCHITECTURE OF HUMAN CYBER-PHYSICAL SYSTEMS – PART I: FUNDAMENTAL CONCEPTS

RA(HCPS)

Prof. Dr. Damm, Werner (Carl von Ossietzky Universität Oldenburg, Germany), Prof. Dr. Hess, David (Vanderbilt University), Prof. Dr. Schweda, Mark (Carl von Ossietzky Universität Oldenburg), Prof. Dr. Sztipanovits, Janos (Vanderbilt University)

Prof. Dr. Bengler, Klaus (Technische Universität München), Biebl, Bianca (Technische Universität München), Prof. Dr. Fränzle, Martin (Carl von Ossietzky Universität Oldenburg, Germany), Dr. Hagemann, Willem (Carl von Ossietzky Universität Oldenburg), Held, Moritz (Carl von Ossietzky Universität Oldenburg), Dr. Ihme, Klas (DLR – Institute of Transportation Systems, Braunschweig), Kacianka, Severin (Technische Universität München), Kerscher, Alyssa J. (Vanderbilt University), Prof. Dr. Lehnhoff, Sebastian (Carl von Ossietzky Universität Oldenburg), Dr. Luedtke, Andreas (DLR – Institut für Systems Engineering für Zukünftige Mobilität, Oldenburg), Prof. Dr. Pretschner, Alexander (Technische Universität München), Dr. Rakow, Astrid (Carl von Ossietzky Universität Oldenburg), Prof. Dr. Rieger, Jochem (Carl von Ossietzky Universität Oldenburg), Prof. Dr. Sonntag, Daniel (Carl von Ossietzky Universität Oldenburg und DFKI-Deutsches Forschungszentrum für Künstliche Intelligenz, Nds.), Dr. Schwammberger, Maike (Carl von Ossietzky Universität Oldenburg), Austel, Benedikt (DLR – Institut für Systems Engineering für Zukünftige Mobilität, Oldenburg), Dr. Unni, Anirudh (Carl von Ossietzky Universität Oldenburg), Dr. Veith, Eric (OFFIS e. V., Oldenburg)

We propose a reference architecture of safety-critical or industry-critical human cyber-physical systems (CPSs) capable of expressing essential classes of system-level interactions between CPS and humans relevant for the societal acceptance of such systems. To reach this quality gate, the expressivity of the model must go beyond classical viewpoints such as operational, functional, and architectural views and views used for safety and security analysis. The model does so by incorporating elements of such systems for mutual introspections in situational awareness, capabilities, and intentions in order to enable a synergetic, trusted relation in the interaction of humans and CPSs, which we see as a prerequisite for their societal acceptance. The reference architecture is represented as a metamodel incorporating conceptual and behavioral semantic aspects. We illustrate the key concepts of the metamodel with examples from cooperative autonomous driving, the operating room of the future, cockpit-tower interaction, and crisis management.

CCS CONCEPTS • Human Centered Computing • Computer Systems Organization • Law, Social and Behavioral Sciences

Additional Keywords and Phrases: real-time systems, cyber-physical systems, architecture, interaction design

1. INTRODUCTION

Today's complex critical systems cannot be designed, built, and maintained without creating virtual models. Following the initial joint initiatives of INCOSE (https://www.incose.org/) and OMG (https://www.omg.org/) in the early seventies to provide industry standard modeling frameworks supporting complex system and system of system development, mutually reinforcing research efforts that were initiated in the EU and the USA now span more than three decades. The overarching goal of these efforts has been extensions to include more viewpoints in software and system design. Key projects addressing this goal in Europe were SPEEDS[1], SPES[2], CESAR[3], CRYSTAL[4], ENABLES3[5], extended to support life-cycle management and product line design (such as CREST[6]) and cooperative decision making (DANSE[7], D3COS[8], HOLIDES[9], AUTOMATE[10]). These projects also support a

1 https://cordis.europa.eu/project/id/033471

2 https://news.safetrans-de.org/ausgabe-2012-02/spes-2020.html

3 https://cordis.europa.eu/project/id/100016

4 https://www.crystal-artemis.eu/

5 https://cordis.europa.eu/project/id/692455

6 https://crest.in.tum.de/

7 https://cordis.europa.eu/project/id/287716

8 https://cordis.europa.eu/project/id/269336

9 https://cordis.europa.eu/project/id/332933

10 https://www.automate-project.eu/

multitude of analysis methods (safety, security, real-time, power, reliability, availability such as in SACRES[11], SPEEDS, ASSUME[12], ARAMIS I, II[13], EMC2[14]). The results have been integrated into industrial design flows (such as AUTOSAR[15] in its many variants, ARP 4754A[16], …). In the USA, the driving force that pushed these trends in the focus of academic and industrial research was a series of DARPA and NSF initiatives in embedded software and systems [58], sensor networks and networked embedded systems, and cyber-physical systems (CPS) [3] [59] [60] [62]. These international research efforts have substantially pushed the state of industrial practice. They motivated the establishment of the Industrial Internet Consortium in the USA (see https://www.iiconsortium.org/) and the Industry 4.0 Initiative in Germany, and they prepared the way to model-based design processes that build on complete virtual twins, which are seen as a prerequisite for on-line adaptability of highly autonomous systems in multiple application domains.

This paper builds on the substantial experience in model-based design by focusing on challenges stemming from the increasing level of autonomy of cyber-physical systems (CPSs) and their deployment in environments that are at best partially known at design time. As key tasks and activities such as perception and decision making are increasingly shifted from humans to CPSs, the abstraction levels of interactions between humans and CPSs increases drastically from low-level control to joint high-level analysis, reasoning, and planning. This shift creates a new level of system design that requires new types of mutual introspection into state, capability, intent, and strategy development between cooperating partners, be it humans or CPSs. It also calls for the seamless communication and interaction of such partners and a quality of dialogue and introspection comparable to trained teams of human experts. It requires achieving a mutual understanding of complementary skills and expertise in jointly assessing and resolving critical situations, such as in emergency rooms or control centers.

To establish such "high-level" interaction, we must also tackle the fundamental challenge of incomplete and/or partial information, particularly in the context of human-machine teams, both among cooperating team members and between the team and its *environment*. The *environment* of an ego system consists of all artefacts of the real world that are *relevant* for allowing the system to achieve its objectives and/or are influenced by the system. The term *relevant* is elaborated below; for the purpose of this introduction, we use it in an intuitive sense: e.g., for an ego-vehicle driving in Paris, the weather conditions prevailing in Seattle are irrelevant in contrast to the weather conditions in Paris. We model such artefacts themselves as systems. Systems in the environment of the ego system can be *dynamic* (such as pedestrians in the environment of the ego car, or environmental systems such as wind or rain impacting both dynamics and perception capabilities of the ego vehicle), or *static* (such as a tunnel on a highway). Systems in the environment of an ego system can thus be humans, physical systems, cyber-physical systems, or human cyber-physical systems.

It is by now common knowledge that lack of confidence in the perception chain of highly autonomous systems operating in only partially known environments is a blocking factor in the deployment strategies for autonomously driving vehicles.[17] This challenge is but one instance of the more fundamental problem of achieving a "sufficiently precise" approximation of *ground truth*—a term well established in the AI community—of the environment of highly autonomous systems and systems for high-level decision support. Without the human in the loop, safety cases must guarantee that the control of hazardous situations is at least as good as that of human operators. Even with humans in the loop, the high-level of interaction that emerges with increasingly delegating tasks to machines can easily cause incorrect decisions with potentially fatal results. For example, in medical decision support systems, the incorrect assessment of a patient's state by automatically generated diagnosis can easily cause incorrect treatments with potentially fatal results, unless experts can understand and, if required, question the basis of such diagnosis. This challenge is but a variation of the above challenge: how do we know that all relevant variables needed for assessing the patient´s health state were entered in the medical decision support systems?

In the context of cooperative decision making, an additional layer of complexity must be tackled: perceptions of one and the same ground truth might easily differ between involved cooperation partners, be they technical systems or humans.

Variations of these problems have been studied for more than thirty years. The following list provides just a few examples:

- The discrepancy between states of the autopilot and what the pilot believes to be the state of the autopilot has led to multiple crashes [21] and triggered the emergence of the discipline of human-centered design [61]. This discipline is based on cognitive theories [53] that explain reasons for such misconceptions and lead to measures to address explicitly the misconceptions.
- Game theory has developed models for decision making and strategy synthesis under incomplete information [32] [54].
- The inherent problems of incongruent knowledge of system states in distributed systems has been captured by concepts such as information forks. These problems lead to undecidability results for analyzing safety of such systems [27].
- The inherent delay between perception of the environment and system action on changes of the environment has led to a control theory for delayed differential systems of equations [2], [12], [26].

---

11 https://cordis.europa.eu/project/id/20897

12 https://itea4.org/project/assume.html

13 https://www.aramis2.com/

14 https://www.artemis-emc2.eu/

15 https://www.autosar.org/

16 https://www.sae.org/standards/content/arp4754a

17 SAE levels 4 and 5

- The issue of the safety and explainability of AI-based components [74] in highly automated control and decision making has recently gained widespread attention. For example, [46] show that completely new approaches to building an AI system may be required to build safe and understandable learning components.
- Numerous research papers address the challenge of understanding the human state and the influence of human state on the capabilities of controlling systems, such as in cases involving automated driving, flight control, drone control, emergency room doctors, and nurses [13], [16], [29], [67], [68], [69].
- Although in the design of CPSs the physical environment is a major factor, in HCPS, the human element ties system design to the social context where these systems operate. Design of HCPS that can be parameterized by the social context and policies is the objective of current research [33], [34], [35], [36], [45], [63].
- The civil avionics domain is requiring quality guarantees of sensor components for certain maneuvers that provide a basis for incremental construction of safety cases for such systems.[18]
- Seeing humans and CPS as partners that provide complementary competences in jointly executing complex tasks has been addressed, e.g., in the AUTOMATE project [22], [47].
- The VDA Leitinitiative[19] on highly automated driving is addressing solutions to the above challenges in the context of highly automated driving.

These research directions are of extreme relevance for assuring trustworthy and safe cooperation of humans and highly autonomous systems. But following such individual strains of research is not enough; they must also be tied together in a holistic framework for mastering the understandability and analyzability of safety critical and/or industry critical complex human-CPS systems. As a very first step, then, addressing this challenge requires building a unifying reference architecture of human cyber-physical systems. This reference architecture can serve as a framework for discussing new challenges, understanding potential approaches for design, and explaining novel phenomena such as the emergent behavior of such cooperatives of humans and intelligent machines. This in turn requires us to revisit the vast experience described above in system modeling from the perspective of the scientific challenges. The following central questions emerge:

- What expressivity is needed to even cast these problems into well-defined research questions on such models?
- How can such models be used for early hazard and risk analysis?
- How can they help to build teams with a mutual awareness of the partners' strengths and weaknesses to handle the problem at hand?
- How can they explain what went wrong, so that such systems can learn from failures and (potentially automatically) evolve to increasingly higher levels of safety and trustworthiness?

We propose the following requirement specification on the definition of a reference architecture for human cyber physical systems:

The reference architecture should be capable of:

- Capturing salient classes of system-level interactions between CPS and professional or semi-professional humans relevant for societal acceptance of such systems;
- Capturing all relevant aspects of physical and societal environments of such systems within a specified operational design domain;
- Capturing, expressing, assessing, and adapting beliefs about physical and societal environments and their confidence levels;
- Reasoning about uncertainties caused by lack of knowledge and observation noise;
- Capturing strategies for self-adaption and self-evolution in dynamically changing contexts;
- Supporting ex-ante, on-line, and ex-post analysis of all incidents/states potentially impacting safety, trust, compliance to ethical and/or societal principles or regulations;
- Supporting seamless cooperation of mixed teams of humans and CPS to reach shared goals within given time-frames on multiple levels of interaction;
- Capturing temporary or persistent formation of coalitions;
- Supporting analysis of realizability of system goals and synthesis of strategies to achieve such goals, possibly involving coalition partners;
- Supporting the assessment of trustworthiness of Human-Cyber Physical Systems (HCPS);
- Supporting justifications for actions chosen by HCPS subsystems;
- Serving as a conceptual framework for building such systems in multiple application domains through specialization.

---

18 RTCA DO-365 Revision B, March 18, 2021 Complete Document MINIMUM OPERATIONAL PERFORMANCE STANDARDS (MOPS) FOR DETECT AND AVOID (DAA) SYSTEMS

19 https://www.vda.de/de/themen/digitalisierung/Autonomes-und-vernetztes-Fahren

This paper proposes a reference architecture for human cyber-physical systems meeting the above criteria, which will be referred to as RA(HCPS), significantly extending earlier research [18]. Section 2 introduces key terms and gives a short introduction to characteristics of four complimentary examples of human-cyber-physical systems used throughout this paper to illustrate and motivate the introduced concepts. In Section 3, we elaborate on the key elements of the reference architecture, and we motivate the selection of its central concepts through the running examples introduced in Section 2. This paper is complemented by two companion papers, [7] and [17], co-submitted to this journal. In Part II [7] we demonstrate the applicability of this reference architecture by refining the abstract view on perception and communication of Section 3 towards requirements on the actual implementation of human-machine interaction. This paper also demonstrates the applicability of these concepts by instantiating the reference architecture in a case study of human-machine interaction in the cockpit. Part III [17] provides a game-theoretic semantic foundation of RA(HCPS), a fundamental prerequisite for any analysis methods, and illustrates applicability of system analysis methods to instantiations of the reference architecture through a preliminary hazard analysis of the cockpit case study.

The conclusion of this paper points to future research directions. An Appendix proposes a shared nomenclature for the moral aspects of HCPS.

2. BASIC CONCEPTS AND TERMINOLOGY

2.1 Aggregation Levels of HCPS

We follow the taxonomy introduced by the SafeTRANS roadmap on Future Man-Machine systems[20] [57] in different aggregation levels of such systems. Unfolding these from bottom to top, we thus use the term "*system*" generically for:

- *individual systems* (e.g., a car, a driver, an operator in a control room, an aircraft),
- *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),
- *homogenous collections of systems* based on networked information exchange (e.g., adaptive routing of traffic flow, air traffic control system) where all systems belong to the same application domain,
- *heterogenous collections of systems* (e.g., smart cities) addressing simultaneously goals of multiple (homogenous or heterogenous subsystems)

This paper addresses safety-critical and/or industry critical Human-Cyber-Physical Systems only. We will use examples from transportation, medical applications, and emergency rescue operations, which are situated at different aggregation levels. The following table summarizes the key complimentary characteristics of the running examples.

Application Name	**Short Description**	**Aggregation Level**	**Characteristics**
Autonomous driving	SAE level 4 vehicles operating autonomously in urban traffic, driver is not required to monitor traffic, but must be able to be reengaged in case of major system failure. Vehicle must thus guarantee high confidence in perception chain to create internal digital model of environment of ego system, must be capable of predicting other traffic participants maneuver, and must select its own maneuvers guaranteeing safety, compliance to traffic regulations, and meeting ethical standards	Individual system	Highly autonomous ego system in highly complex environment with comprising both humans and other systems Not yet available in market. A key challenge in market introduction rests in having mixed traffic, where most other vehicles are operated by humans. Deployment contexts differ regarding support in decision making by infrastructure.
Operating Room of the Future	Teams of surgeons, anesthesiologists, registered nurses are supported in the operating room by multiple medical devices ranging from monitoring all relevant parameters of patient status to creating virtual reality-based guidance for operating procedures to highly automated robot assisted surgery technics	Groups of systems	Local communication among multiple humans with different qualifications working as a team with multiple cyber physical system
Cockpit-Tower interaction	Cockpit crew coordinates with tower and assisted with multiple cyber physical systems ranging from monitoring all relevant aircraft health state parameters, sensor systems, autopilot systems, and radio communication system to select safe flight trajectory of aircraft	Homogenous system of system	Highly trained humans interacting with local and remote humans and cyber-physical systems
Emergency Response	Stakeholders in emergency response scenarios are military organizations, international organizations, government and non-government organizations and possibly private organizations. The heterogeneity of stakeholders with different structures and characteristics, and the use of diverse legacy equipment with different levels of complexity are increasing the complexity of Emergency Response operations. The response to a crisis is the result of the activities of different services (e.g. police, medical care, rescue forces, and fire-fighting), interacting vertically (i.e. with components of the same organization) and horizontally (i.e. with components of other organizations), in a complex environment.	Heterogenous system of system	Multiple organizations with different competences and different roles coordinating activities of their capabilities within a city

20 https://www.safetrans-de.org/de/Aktivitaeten/Roadmapping.php

2.2 Terminology and Acronyms

For ease of reference, this section summarizes key terms and acronyms used in this paper.

Action	Synonym for actuation capability.
Belief	Each layer maintains a representation of its beliefs about the environment of the ego system, as observed through its perception system. Both perception and belief formation are influenced by the current state of the ego system. The term "beliefs" can refer to descriptive beliefs ("is" states) and prescriptive beliefs ("ought to be true" states). For clarity, we refer to each type of belief as descriptive or prescriptive.
Capability	The services a system can invoke in order to achieve its goals. Capabilities include communication, coordination, delegation, perception, and actuation (which refers to controlling movement of the ego system, including non-functional characteristics as well as key performance indicators on such movements). We use the term "action" as synonym for "actuation capability."
Ego system	The system under discussion; the attribute "ego" is used to emphasize that all statements about the system are made from the perspective of this particular system.
Environment	The environment of an ego system consists of all artefacts of the real world, which are relevant for allowing the system to achieve its objectives and/or which are influenced by the system.
Ethics	"Ethics/ethical" refers to the theoretical reflection and discussion of morality/moral values and norms with regard to their validity, acceptability and legitimacy. Sometimes the results of ethical reflections and discussions are expressed in ethical codes or systems.
Descriptive states	These states are actual situations about the self or the world, or they are predictions about how situations are likely to change given current information or planned actions. (Other terms are also used. Some use the term "positive" in contrast with "normative.")
Ground truth	The state of the environment of the ego system as seen by an omniscient observer.
Goal	The requirement specification of an ego system defines for each of its layers the goals the system is to achieve. Goals are typically (partially) ordered based on their priority. We use the terms "goals" and "objectives" as synonyms.
Health state	The health state of a CPS is determined both by the collection of current failures of the system as well as by the detection of violations of system integrity such as through cyber-attacks. For humans, this state includes any medical or mental conditions that can impair or even block the ability to perform any of the functions of an ego system.
Prescriptive state	The term "prescriptive" is used to describe states in human-cyber-physical systems that refer to "ought" conditions or what states of a situation or scenario "should" be.
Layered architecture	A decomposition of a system's overall functionality into linearly ordered abstraction layers, each of which focuses on particular aspects of the functionality and assumes services offered by lower levels as given. The layers abstract from the details of their realization.
Moral	"Morality/moral" refers to a set of commonly accepted (evaluative or normative) standards (e.g., traditions, custom, and conventions) that can vary between different places and historical times.
Norms	"Norms" are prescriptive rules about what is acceptable in relations among individuals, groups, organizations, or other social units. Norms can be formalized as law, standards, or codes of behavior and backed up by institutionalized systems of rewards and sanctions.
State	Ego-systems have a state, which influences all layers of the ego system and is influenced by all layers of the ego system. For human ego-systems, the ego-state comprises all aspects of human states that impact their interaction with other systems, such as state of short-term memory, frustration, alertness, etc. For technical ego-systems, the ego state subsumes the valuation of all its system variables.
System objectives	We use the term "objective" as a synonym for "goal"; see entry for "goal."
Values	Values are general principles that guide action. Values are about what is important, good, beneficial, or desirable.
Virtual twins	A complete digital copy of the system that represents all layers of the system with a degree of accuracy in matching the capabilities of these layers as ultimately determined by its safety requirements.

AUTOSAR	AUTOSAR (AUTomotive Open System ARchitecture) is a global partnership of leading companies in the automotive and software industry. The partnership develops and establishes the standardized software framework and open E/E system architecture for intelligent mobility. See https://www.autosar.org/
SARP	The Aviation Safety Management and Recommended Practices SARPs are intended to assist states in managing aviation safety risks, in coordination with their service providers. Given the increasing complexity of the global air transportation system and the interrelated aviation activities required to assure the safe operation of aircraft, the safety management provisions support the continued evolution of a proactive strategy to improve safety performance. See https://www.icao.int/safety/safetymanagement/pages/sarps.aspx
ARP4754A	Guidelines for Development of Civil Aircraft and Systems, see https://www.sae.org/standards/content/arp4754a/
INCOSE	The International Council on Systems Engineering (INCOSE) is a not-for-profit membership organization founded in 1990 to develop and spread the interdisciplinary principles and practices that enable successful systems. See https://www.incose.org/.
Industrie 4.0	The I40 strategy aims to ensure an industry fit for future manufacturing in Germany. It supports the integration of cyber physical systems (CPS) and Internet of Things and Services (IoTS) with an eye to enhance productivity, efficiency and flexibility of production processes and thus economic growth. See https://www.bmwk.de/Redaktion/EN/Dossier/industrie-40.html.
IoT	Internet of Things; see https://www.iiconsortium.org/.
OMG	The Object Management Group® Standards Development Organization (OMG® SDO) is an international (27 countries), membership-driven (230+ organizations) and not-for-profit consortium. See https://www.omg.org/.
SAE	SAE is a global association of more than 128,000 engineers and related technical experts in the aerospace, automotive, and commercial vehicle industries. The association creates standards for such systems. See https://www.sae.org/standards.
SAE level 4	At Level 4, vehicles are completely responsible for all driving and navigational tasks, and thus are called autonomous. These self-driving cars can autonomously transport passengers who do not need to be engaged or ready to take control of the vehicle (See https://www.sae.org/blog/sae-j3016-update). Some authors prefer the term *automated vehicles*, notably if these are equipped with wireless communication capabilities allowing exchange of perceptions and coordinated maneuvers, then abbreviated as CAV (connected and automated vehicles). Since such capabilities are not mandated for Level 4 vehicles, we use throughout this paper the term *autonomous vehicles*, highlighting their key capability of driving without driver interaction.
VDA	Verband der Automobilindustrie, the German Association of Automotive Manufacturers. See https://www.vda.de/en.
VDA Leitiniitative	A suite or projects coordinated by the German Association of Automotive Manufactures to support highly autonomous driving.

2.3 Guide to the structure of the presentation of the proposed Reference Architecture for Human Cyber Physical Systems

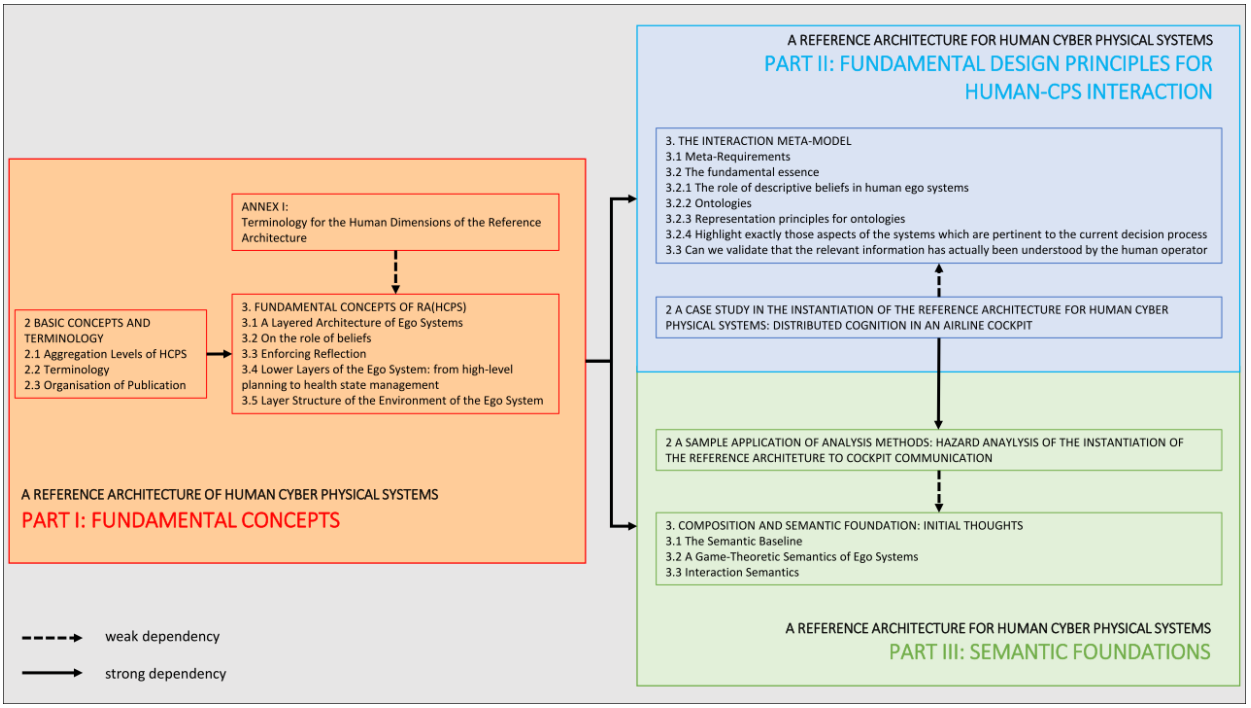

Figure 1: Overview of Structure and Flow Reference Architecture for Human Cyber Physical Systems

We have structured the presentation of the Reference Architecture for Human Cyber Physical Systems in three parts:

- Part I describes its fundamental concepts, which we see as mandatory ingredients in any such system design. It uses four classes of running examples throughout the text, which are situated at different aggregation levels of such systems, as introduced in Section 2.1. The fundamental barrier of both humans and cyber-physical systems to be limited by their perception capabilities and their own interpretation of such perceptions motivates the need to anchor *beliefs* (about other systems and it´s environment) as first-class citizens. This calls as well for the constant need of *reflection*, both on plausibility of beliefs, as well as on observing moral and ethical principles, which must govern the increasingly automated actions of such systems. This part completely abstracts from the realization of the interaction between Humans and Cyber Physical Systems, and only explains the semantics of such systems intuitively.
- Part II focuses on fundamental design principles for Human-CPS interaction. To understand the type of challenges, it´s Section 2 demonstrates the multiple pitfalls and established solutions of such interaction designs using the interaction within the cockpit of a civil aircraft and in the communication with the tower. Part II builds heavily on concepts of cognitive psychology in order to answer the question of what aspects of the controlled system and it´s environment must and can be communicated under real-time constraints leading to actual perception and correct interpretation of such perceptions by human actors, taking into account the professional training mandated for such operators in safety-critical and/or industry critical systems control, such as in controlling a civil aircraft.
- Part III complements these by defining a rigorous game theoretic semantics of human-cyber-physical systems. The fundamental barrier of both humans and cyber-physical systems to be limited by their perception capabilities and their own interpretation of such perceptions demands the conception of a new category of games, which form strategies based on their beliefs about the arena of the games. The safety risks stemming from this are illustrated in it´s Section 2; it shows how by anchoring beliefs as part of the reference architecture classical safety analysis methods are enforcing detection of the risks of such misconceptions and their criticality.

Dependencies are indicated by strong and weak arrows, weak arrows indicating, that it is beneficial but not strictly necessary to read the text in this order. In particular, up to the shared cockpit case study, Parts II and III can be read completely independently.

3. FUNDAMENTAL CONCEPTS OF RA(HCPS)

This section captures what we consider to be the common essence of critical HCPSs. We capture the common essence in one meta-model, which we refer to as the reference architecture of HCPSs – RA(HCPS).

3.1 A Layered Architecture of Ego Systems

We describe this reference architecture from the perspective of a given singular ego-system, which can be either a CPS or a human (interacting with a CPS), a group of HCPSs, or an organization that is situated at any level in the aggregation hierarchy as described in Section 2.1. Please note that this is non-standard usage of the term “ego system.” We will elaborate below on how the proposed reference architecture for ego-systems naturally specializes to this broad spectrum of possible concretizations.

A fundamental aspect of our model is induced from the inherent challenge of imperfect information. To be able to express (such as in meta-requirements for system design) requirements on the ego´s perception system, we associate with each system two fundamental views:

- *An ego-system view* comprising all aspects of system organization and state that are *known* to the system, depicted below in Figure 1 in blue;
- *An environment view* of the ego system comprising the *ground truth* of all relevant artefacts of the environment of the ego system, depicted below in Figure 1 in red.

The distinction between *beliefs* about the environment (as formed by the ego-system) and its *ground truth* (as seen by an *omniscient observer*) is fundamental to our modelling approach. No system can achieve perfect observation of its environment. A universal meta-requirement for system design is to guarantee that beliefs are always *sufficiently precise* approximations of ground truth. (The semantic foundation of RA(HCPS) provided in our companion paper [17] makes it possible to give a formal definition of this meta-requirement on system design.)

Figure 1 below proposes a layered structure for the reference architecture of an ego system.

- We assume ego-systems share a hierarchical organization of subsystems[21] that can be structured into eight *layers*, as shown on the left-hand side of Figure 1, color coded blue.
- Each subsystem has a *control* component, which determines the currently active objectives pursued by the system (the kind of objectives considered are dependent on the aggregation level of the system).
- Ego-systems have a *state*, which influences all layers of the ego system and is influenced by all layers of the ego system. For human ego-systems, the ego-state comprises all aspects of human states that impact his/her interaction with other systems, such as state of short-term memory, frustration, alertness, etc. For technical ego-systems, the ego state subsumes the valuation of all its system variables. We distinguish these state components from those which are uncontrollable by the ego system, such as through illness, degradation or even failure of (sub-)systems possibly caused by cyber-attacks.
- At each point in time, an ego system strives to achieve a set of *goals*. Goals vary depending on the aggregation level of the ego system and on the level of the layer in the hierarchical models of the ego system. We assume that goals are partially ordered and that the order relation reflects the priority with which goals are pursued.
- At each point in time, an ego-system has a set of *capabilities* for achieving its goals. Capabilities vary depending on the aggregation level, the level of the layer in the hierarchical model, its *health state*, as well as environmental conditions influencing physical laws. Capabilities include communication, coordination, delegation, perception, and actuation (which refers to controlling movement of the ego system, including non-functional characteristics as well as key performance indicators on such movements).
- The *health state* of a CPS is determined both by the collection of current failures of the system as well as by the detection of violations of system integrity such as through cyber-attacks. For humans, this includes any medical or mental conditions impairing or even blocking the ability to perform any of the functions of an ego system.
- Ego-systems can assume different *roles*, which form a particular class of states. They impact the current set of goals pursued by the ego system, as well as its available capabilities. Role changes can be a deliberate act of the system or caused by external events. Changing roles influences all layers of the ego system.
- At each layer, the *control* component determines how priorities of goals can be changed dynamically, either deliberately by the ego-system in assuming a different role, or through escalation from perceptions and subsequent control actions of lower levels, or through interaction with other ego-systems. The control component also chooses to activate or deactivate layer and role specific capabilities.
- Each layer maintains a representation of its *beliefs* about the environment of the ego system, as observed through its perception system. Both perception as well as belief formation is influenced by the current state of the ego system. The term “beliefs” can refer to *descriptive beliefs* (“is” states) and *prescriptive beliefs* (“ought to be true” states). For clarity, we refer to each type of belief as descriptive or prescriptive. (See the appendix “Terminology for Human Dimensions of the Metamodel.”) We discuss the key role of beliefs in Section 3.3 below.

  Examples of ego systems are cars, aircraft, operators of control rooms, surgeons, drivers, and hospitals.

The right-hand side of Figure 1 provides a view of the ground truth of all relevant artefacts of the ego system´s environment as seen by an omniscient observer. We refer to this view as the *ground truth view of ego´s environment*, and denote it by *ego(ENV)*. It subsumes all aspects of the ego-systems environment, whose ground truth is *relevant* for the ego system. The precise meaning of relevance is defined in a companion paper [15]. For the purpose of the current paper, this term is given an intuitive interpretation: an artefact is relevant to the ego-system if it can influence the system in a way that could prevent it from achieving its goals. A meta-requirement for any HCPS design can be defined by comparing the ego-system´s current descriptive beliefs about its environment with the current ground truth as represented in *ego(ENV)* as follows: for all relevant artefacts in *ego(ENV)*, its descriptive belief about this artefact and the actual ground truth should, after a fixed bounded goal dependent latency, differ at most by some goal dependent margin from its ground truth. The omniscient observer could also have prescriptive beliefs—such as some "reference expression" of social norms as codified for everyone—that may be different from the individuals understanding of it.

[21] Strictly speaking, this is a *functional* decomposition into those subfunctions, which we consider essential for HCPS. The actual implementation may of course use a different *logical architecture* in distributing these functions into deployable units on the *target architecture.*

The environment is constituted by multiple systems listed in Figure 1 as S1, S2, …, Sn, …, which all comply with the same reference architecture. Hence, systems can be composed hierarchically with an unrestricted depth.

Examples of instances of systems in *ego(car)(ENV)* are the driver of the car, traffic participants in its surrounding environment, their relative distance and velocity, the state of the traffic light ahead, and the conditions of the road surface. The following are also included: descriptive beliefs about their states, plans, and associated goals. For example, "the driver in the car ahead is drunken and cannot control the car," or "the pedestrian on the sidewalk is about to cross the street."

The information flow shown in red reflects that the ego system is attempting to observe the ground truth of relevant systems in its environment through its perception sub-system. These perceptions lead to the formation of descriptive beliefs about the ground truth of perceived systems, often referred to as the *world model* of the ego system. This world model forms the basis for the decision making of the ego system, such as the invocation of capabilities of the system. Actions of the ego-system impact the ground truth of its environment, as shown by blue arrows.

For technical ego systems with enough computational resources, all layers of the ego- system will be active concurrently. Humans, however, are constrained by the number of cognitive tasks they can pursue concurrently. For humans, we allocate in the reflex layer all activities that are performed subconsciously and assume that the reflex layer as well as the moral system layer are always active.

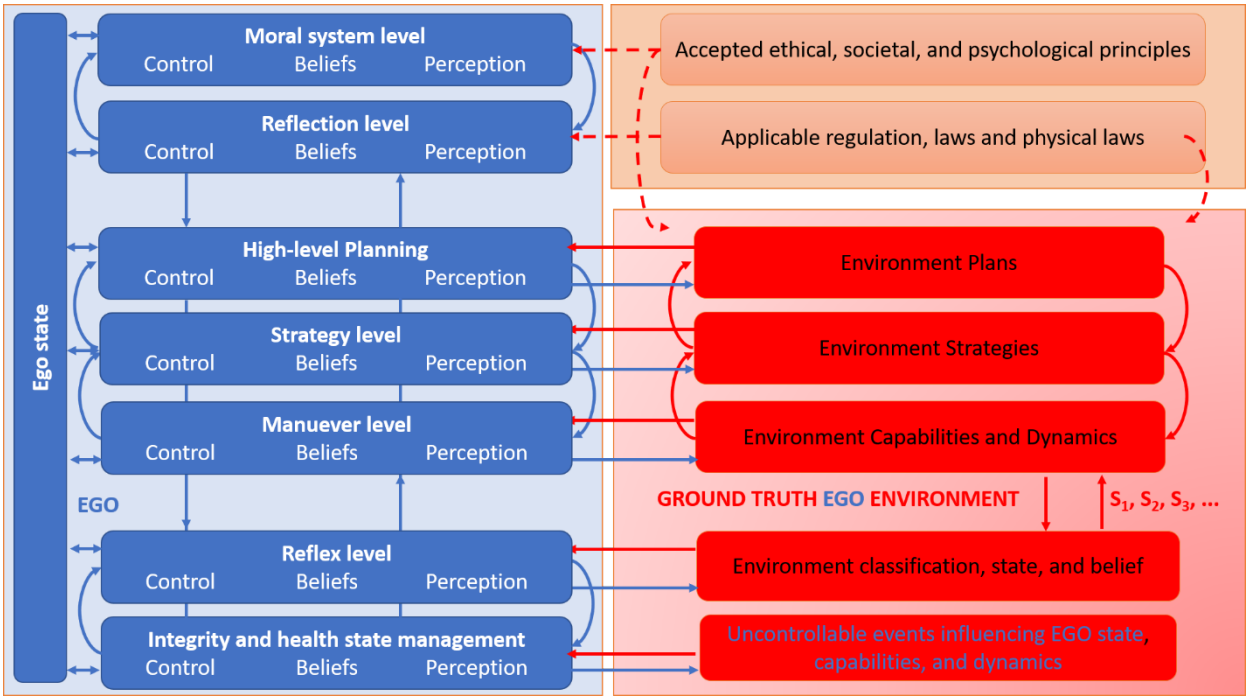

Figure 2: Constituents and Layers of the reference architecture for human cyber-physical systems

Before we further elaborate on the layers of our meta-model, let us introduce "its spirit" by way of two simplified examples.

Example

Figure 2 below shows components of subsystems of autonomous cars and their current state. The car has identity c, its current speed is 60 mph, and it knows other cars with identities c2, c3, c1 as well as a traffic control center HwyCTR_h2. By entering its scope, it signed electronically a contract that gives the highway control center the authority to temporarily change goals of car c. Car c coordinates its behavior with car c1. Car c communicates with other systems such as c2 through communication channels. Car c perceives other cars in its proximity based on raw data of its sensor systems (formally input signals driven by the environment of S) and its subsystems for classification of environment artefacts leading to environment descriptive beliefs, and Car c also controls its longitudinal and lateral acceleration based on these descriptive beliefs. In the role "Highway," the car follows the four goals shown (with that order of priorities). To this end, it invokes capabilities 1-8, out of which the first six are active. Availability of capabilities depends not only on the role but also on the health state of the systems (e.g., is the video camera defect? is car2infrastructure communication available?). Health state "all" indicates that all relevant subsystems of car c are currently available. The available capabilities also depend on the environment of the system (e.g., whether the road is icy). The figure also highlights the descriptive beliefs of car c about its environment, relating to position, speed, and distance of vehicles in its proximity, as well as environment conditions, as indicated by environment mode sunny. Note that beliefs of c about its environment differ from ground truth: some light rain has not yet been detected, the distance between c1 and c was overestimated, and the velocity of c1 was underestimated.

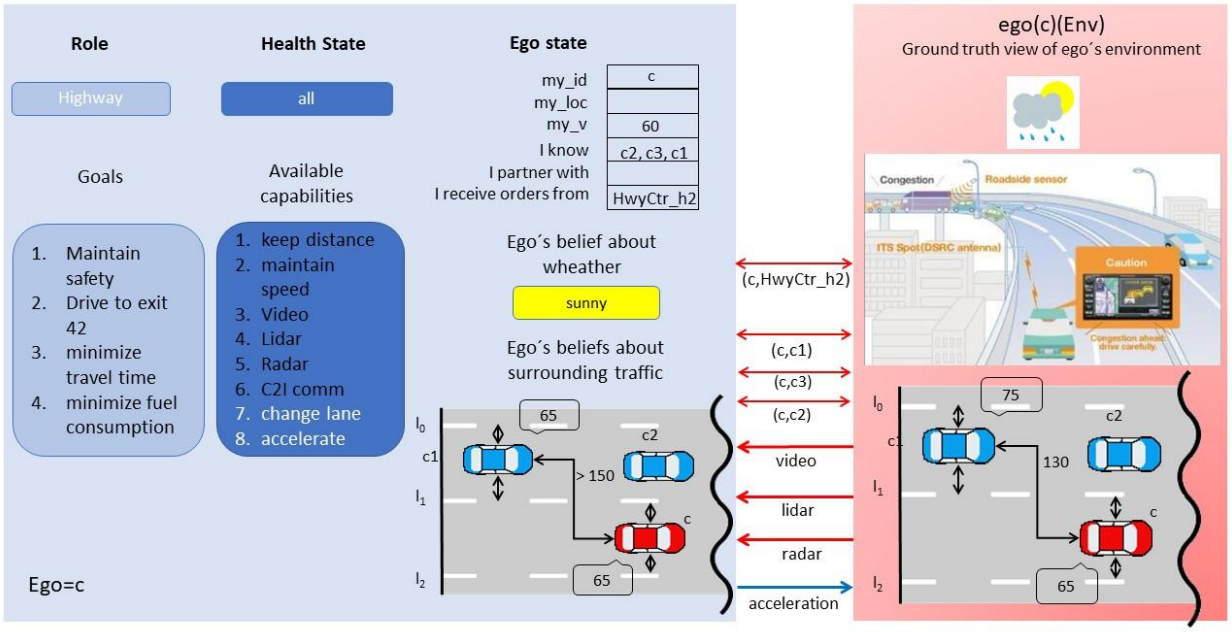

Figure 3: A sample instance of an ego system for highly automated driving

Example

> Jane (a physician) has just started a routine surgery on patient P in the operating room, when she receives an emergency call to treat a patient with a complicated heart failure, who is just being taken to a second operating room. As the most experienced doctor on the team for heart surgeries, Jane hands over control to continue the surgery of patient P to Andy (another physician), who now controls the team and gets status updates from both the attending physician, the anesthesiologist, and the nurse attending patient P. In our model, the emergency call to Jane is handled by Jane´s "internal" control at the high-level planning level by (1) delegating the role of being the head surgeon of P to Andy, (2) accepting the new role of being the head surgeon in charge of the heart-emergency patient, and (3) switching the operating room. Her role as lead surgeon on P allows her to delegate authority of becoming the lead surgeon for P to Andy. This new role for Andy gives him privileged access to all information regarding the status of patient P.
>
> Andy is done with this surgery after two hours, just slightly behind his official end of working hours for the day, and he changes clothes and is ready to drive home. This change of roles is reflected in our meta-model by a deliberate role change from "doctor" to "driver." In other words, the moral system pertinent to his role as a doctor is now replaced by a moral-system that he shares with all drivers in his home country, in particular regarding traffic regulations. With this role change, he loses the right to access the status information of patient P. As he drives home, he passes a scene where a traffic accident just occurred: a car has crashed into a drunken pedestrian P´, who has simply ignored all traffic rules and stepped out into the heavy traffic. Andy immediately parks his car and performs emergency rescue operations. In our model, when P´ became intoxicated (this being part of his current ego-state), the control component of P´ at its moral system layer has temporarily deactivated his moral system at least partially, in that traffic rules were completely ignored. Andy is changing his role to become doctor, and he performs emergency treatment on P´.

Returning again to the right-hand side of Figure 1, the upper two components of the environment of the ego system are of virtual nature only and are depicted in pink. They are referred to as "prescriptive states" in the Appendix "Terminology for Human Dimensions of the Metamodel" and represent the following:

ENV1 the moral system guiding the actions of the ego system; and

ENV2 specific regulatory, legal, and ethical code constraints on the ego system.

For the purpose of this paper, ENV1 and ENV2 are considered to be stable for the current deployment context of the ego system – in other words: they are considered to be location dependent and time independent.

Examples

- Different emissions standards for a car deployed in the US has to meet different emission standards than in Germany, and regulations regarding type certification for autonomously driving cars differ between US and Europe.
- There are different ethical principles between the US and Germany associated with soft laws and ethics guidelines for decision making in levels of hospital-based, health-care provision for elderly people.

- Ethical principles for algorithmic conflict resolution in non-avoidable crash situations of autonomous systems differ between US and Germany.

The assumption that such principles are time invariant is of course a gross oversimplification. The technology push of increasingly deploying autonomous and/or self-learning systems has triggered debates leading to the development of new principles for AI design [5, 4, 6, 10, 11, 23, 24, 28, 30, 31, 38, 39, 41, 43, 44, 49, 55, 56, 66, 70, 73] and new types of certification procedures [8, 9, 14, 19, 20, 37, 38, 42, 48, 51, 52, 64, 71, 72]**.** The only technical assumption we make is that the dynamics of such changes is much slower than the dynamics of the ego systems and other environment components depicted in red in Figure 1.

The layered architecture of the ego view extends classical SENSE – PLAN – ACT models by anchoring *beliefs* as first class citizens on all layers of the reference architecture, and enforcing *reflection* on both the plausibility of beliefs as well as on the impact of the actions chosen by the ego system on other systems.

3.2 On the fundamental role of beliefs

We recognize the fact that ego-systems are inherently incapable of having perfect observations of the ground truth of their environment, by anchoring the concept of descriptive beliefs as a first-class principle in our reference architecture at all layers of the meta-model. In other words: ego-systems are moving in often highly complex deployment contexts, which are only partially observable by imperfect means coming with inherent distortions, and thus can only form descriptive beliefs or what is sometimes called "mental models" or "world models" of their deployment context. Sources of distortions are:

- Limited direct observability:
    - A constituent system has by itself only limited direct access to information, i.e., all information is "perceived" directly through its sensory systems (whether technical or human).

  Example

    - The selection and physical arrangements of sensor systems in highly autonomous vehicles determines which aspects of the environment are inherently not observable, e.g., because they are out of range of the sensor system.
- Distorted perception:
    - For technical systems, even for directly observable real-world entities, the perception of reality will differ from reality due to effects ranging from noise at the physical level of sensors to inherent trade-offs in excluding false negatives versus guaranteeing high detection rates in object classifiers deployed in later stages of the perception chain.

  Example

    - Radar systems produce "ghost objects" through reflections of radar waves, e.g., from walls in a tunnel.
    - Video systems in highly autonomous cars are known to miss detection of pedestrians in front of the car due to lighting conditions, or to interpret an image of a human which is part of an advertisement of a truck to be a human.
    - Use of inadequate world models used for the interpretation of observations: even with perfect sensory systems, actions of the system will fail drastically if perceptions are interpreted in an inadequate world model.

    - For humans, perception can be distorted because of lack of attention, limited processing capabilities, the influence of human states such as stress and fatigue, or simply the lack of knowledge of interpreting the perceived artifacts or situation correctly. The following are also sources of distortion:
        - 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.
    - Limited perception bandwidth: perception is inherently limited by the perception bandwidth
    - Failures: These effects are aggregated by failures in communication and sensory system, which, without proper error-detection and recovery mechanisms, can lead to arbitrary large gaps between perceptions of reality and reality itself.
    - Lack of awareness of the health state of the perception system. The health state of a system is uncontrollable; hence, as part of *ego(ENV)*, it is not directly observable. Thus, incorrect descriptive beliefs about perception components may lead to distorted perception.

- Distorted communication:
    - Whenever sufficient situational awareness can only be gained by communication, it is subject to distortion in the time domain. The inherent physically distributed nature of HCPS applications makes it difficult to guarantee sufficient bounds on communication jitter. Merging such information with directly perceived information can lead to descriptive beliefs about the environment that are both distorted in the space and time domains, even in the absence of failures in sensory- and communication systems.

- Compromised information:
    - All information channels (sensory and communication) can be compromised and thus can be subject to intentional disinformation.

  Examples

  - The autonomous car does not know the conditions under which its radar sensor is distorted and thus believes that an object is ahead of the car, though it is only a reflection.
  - Firefighters in a building can only perceive the immediate local surroundings and must rely on information provided through wireless communication to obtain a more complete situational awareness.
  - An aircraft world model that may not "know" about volcanic ash may falsely classify a cloud of volcanic ash as hail and thus not consider it mandatory to initiate a fly-around maneuver in an already highly delayed flight.

Humans deal with such situations routinely by creating hypotheses about the state of the environment based on previous experience and whatever limited observation is given. Crucial for human performance in dynamic environments is the ability to make predictions and to compare the predictions based on descriptive beliefs derived from the sensory data, a capability also owned by technical systems and anchored in the reflection layer in our reference architecture. A mismatch between prediction and observations can trigger a re-evaluation of descriptive beliefs. Human decisions are often based on predictions, as humans are typically too slow in information processing as well as in implementing actions. This process can run without volition (e.g., in spoken conversations), voluntarily (e.g., when entering a crossing and scrutinizing the trajectories of other egos), or even subconsciously in some highly automated tasks (like shifting gears or steering through a curve). In our reference architecture, we represent all such hypothesis about the deployment context as descriptive beliefs, including descriptive beliefs about dynamics of entities in the environment of the ego-systems.

Examples

- We do not necessarily know the true cause for the sudden collapse of the patient, and yet we will initiate treatments that we expect to be able to stabilize the patient.
- We do now know the exact nature of the road surface in the curve ahead, and yet we want to stabilize the car when passing the curve.

Descriptive beliefs of humans comprise subjective presumptions of the inaccessible ground truth about their environment. Although the ground truth would, for example, assign a real definition to items like "speed," a descriptive belief on speed will hardly do so. It can assign anything from "absolutely don't know" to "probably higher than 100 km/h"; in other words, it ranges over a different mathematical domain. The difference stems from varying confidence levels in descriptive beliefs, as stated below, but it also is structural. Furthermore, descriptive beliefs not only pertain to states, but also to dynamics, as follow:

Example

- "Though the alter car could intrude into my safety envelope, from this situation on, it normally proceeds only until a stop at the line of sight (and I thus believe it will do so in this case)."

Descriptive beliefs are equipped with confidence levels.

Examples

- In differential diagnosis, the doctor not only takes into account one single diagnosis, unless they are "100% sure," but will also take into account less likely causes for the observed symptoms.
- In automated driving, the location of the sun at dawn or dusk may make it impossible to be sure of the type of the vehicle in the opposite lane; however, based on its speed, the car believes with medium confidence that it is a truck.

If multiple stochastically independent observations support the hypothesis expressed in a descriptive belief about the ground truth of an artefact in the environment, a system can justify associating a higher confidence level with this descriptive belief. In general, we assume that confidence levels of descriptive beliefs are annotated with justifications. For technical systems, such justifications are created "bottom-up" along the perception chain. These justifications reflect the plausibility of descriptive beliefs, as discussed above, using redundancy coming from multiple sensor systems, time-series analysis, consistency with physical laws or in general learned behavioral models.

Examples

- An autonomous vehicle will compute with what confidence it believes that the object detected ahead of the car is a person and check that this assessment has been consistently confirmed by analysis of radar and video image streams over the last second

- A crisis management system will create its initial response for addressing a crisis event based on its descriptive beliefs learned from training or previous experience about the capabilities of its subsystems to handle this particular kind of crisis. Its confidence to a subsystems capability of handling a certain type of event will grow from repeated experiences.
- An anesthesiologist will react to a sudden collapse of the patient based on her descriptive beliefs about the cause of the collapse and the anticipated success of the planned intervention.
- A human operator in a missile defense system will use plausibility checks regarding the classification of a potential attack based on intelligence information about likelihood of attacks.

To be able to make reliable prediction of the future behavior of other systems in the environment of the ego system, it is necessary to have the capability to reason about descriptive beliefs of descriptive beliefs. In general, our model has to support higher order introspection; i.e., it must be able to reason about that A believes that B believes that….and so on. This concept is central to the theory of mind [65] as established in cognitive psychology, and in the context of this paper it refers to the need to have introspection in all aspects and states of external systems as far as they are relevant for the ego system. Cooperation between such systems typically requires achieving a *shared situational awareness*. In other words, cooperation requires a consistent understanding about relevant states of neighboring systems that is achieved, for example, by sharing such descriptive beliefs and their confidence levels, which in turn can lead to descriptive belief revisions such that partners "essentially" agree on relevant states.

Examples

- Firefighter A may deliberately decide to leave the current floor of the building, because Firefighter A believes that Firefighter B is taking on the task of covering the entire floor when looking for injured persons. Firefighter A may have based his or her descriptive belief on pre-defined strategies, on how the fire brigade will search the building, or on information received by a wireless channel.
- We illustrate the concepts using Figure 3. In Figure 3, the Emergency Rescue center at Hospital B has an accurate view about the location of the explosion. However, it incorrectly believes that Police Station A has helicopters available, due to a process failure in updating its location information. Since A is closest to the location of the explosion, B believes that A will deploy its helicopters to the emergency site. A´s communication system was tampered with so that it transmits an erroneous location of the site of the explosion. Hence A believes emergency rescue center at Hospital B and the nearby police station to be closer to the site of explosion. Based on this belief, A makes a decision on the assumption that this police station will deploy its rescue helicopter, and that Hospital B will deploy its rescue vehicles to the crisis site.

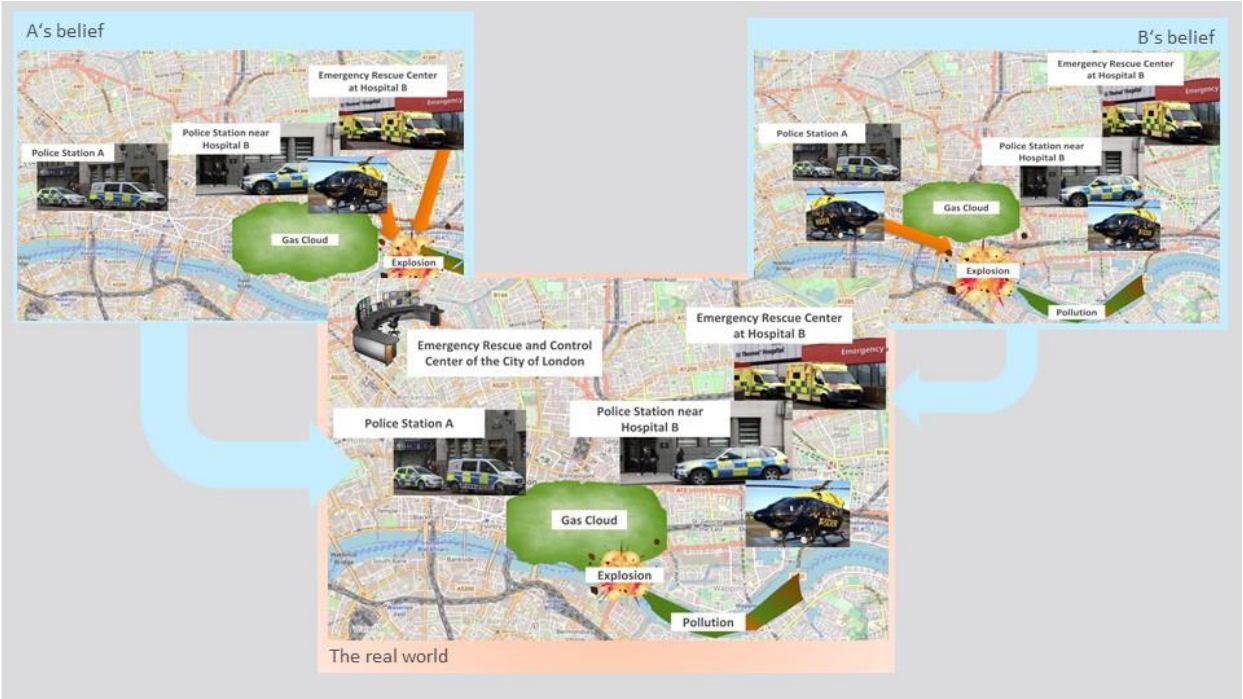

Figure 4: Discrepancy between A´s descriptive beliefs and the ground truth and B´s descriptive beliefs about the location of the explosion may cause initiation of conflicting rescue measures

We extend the set of descriptive beliefs to include beliefs of the environment dynamics. Each constituent system will use its sensors and a comparison of its own prediction of future state evolution with the actually observed future states to adjust its belief about the current environment model dynamics. As for other descriptive beliefs, we will associate a confidence level with such beliefs about the environment´s dynamics.

3.3 Enforcing reflection

Given the increasing degree of autonomy of deployed systems, we consider it mandatory to anchor reflections of the action of the ego system on other systems as a required component of such systems.[22] Principled AI design (see e.g. [24, 25, 39, 40, 41, 42, 48]) calls for such systems never to endanger safety of other systems and to obey principles of fairness, justice, privacy, etc. Thus ego-systems are required to anticipate the impact of their own actions on others (a descriptive belief). But choices for current actions also require internal prescriptive states or moral beliefs, which we define as states in human-cyber-physical systems that refer to "ought" conditions or what states of a situation or scenario "should" be. Goals are situation-dependent prescriptive states that the system wants to reach, or that human actors have in specific situations, and as such are descriptive in nature. However, the process of selecting and/or prioritizing goals calls for a reflection of the compliance of selecting such goals and their priorities with the moral system. For example, when selecting the goal of controlling the aircraft in such a way that it collides with the twin towers, the goal as such describes the state the terrorists want the aircraft to reach, and the terrorists deliberately ignored the moral system in selecting this goal, because they wanted the aircraft to crash.

Reflections on plausibility about beliefs is fundamental in system design. For example, hypotheses about states of environment systems may have to be discarded when new sensor readings are available, such as by changing a descriptive belief about the diagnosis when new symptoms of a patient are observed. Technical systems may maintain multiple hypotheses simultaneously and act according to what they believe to be the most plausible descriptive beliefs about the ground truth of *ego(ENV)*. We do allow for co-existence of multiple possible world models that correspond to multiple hypotheses, where each such world model represents consistent sets of descriptive beliefs. However, observations in different world models can be inconsistent, and only future observations will allow decisions about which of these hypotheses will be maintained. Humans tend to operate mostly based on one-world model, and they consider switching to a model that better fits the current data when predictions fail. In psychology we call this *flexibility*, and human cognition needs a good balance between fixation on a strategy and flexibility. In contrast, during scientific exploration, multiple hypotheses are maintained and tested until enough evidence has been created to have high confidence in a one world-model.

Much like humans, highly autonomous systems will continuously assess the plausibility of such descriptive beliefs, for example, by relying on its descriptive beliefs about prevailing principles and laws.

Examples

- Time-series analysis of video data tracking the environment of an autonomously driving car will consider a snapshot implausible if an object detected in front of the car is first classified as a woman, and then as a bicycle, and then as a shopping bag, because the applicable physical laws disallow such metamorphosis of objects.
- Humans may suffer a similar problem but would interpret this rapid metamorphosis as uncertainty of their descriptive beliefs about the nature of the object, and they would change their strategy to a more cautious approach.
- Human power-grid operators may use a simplified physical model of the grid based on one type of parameter to characterize the current grid state, but if they realize that the control actions are not well predicted by the first model, they may switch to another parameter that reflects a different physical replacement model.
- In the case of a driver entering a crossroad who has previously identified a car that had to yield according to the regulations, the driver first assumes that the other vehicle took notice and acts according to the regulations. Upon realizing that this descriptive belief is wrong, the driver will change plans to avoid a collision. However, some would only honk the horn, depending on the affective/motivational state. Others may even forget to check the other vehicle happily believing that everything is fine until the other person crashes into the driver's car.

In summary, all kinds of descriptive beliefs undergo a continuous process of revisions and updates. Given enough resources, technical systems can maintain multiple possible worlds, and make their actual moves dependent on predictions of evolutions taking into account all maintained possible worlds. With the limited capacity of humans, this approach is infeasible: humans would typically work with one "best fit" model and then switch to a different world model believed to be best fit when the current model is inconsistent with new perceptions.

To be able to act in partially unknown complex contexts and yet choose actions that are not only capable of supporting the ego-system´s current goals, but also to maintain consistency with the moral system, we assume the following layers in the ego-view:

EGO 1 Moral System Layer

The term "morality/moral" refers to a set of commonly accepted (evaluative or normative) standards (e.g., traditions, custom, and conventions) that can vary between different places and historical times. Values are general principles that guide action. Values are about what is important, good, beneficial, or desirable. In some societies and some formal systems, values may be organized hierarchically. "Norms" are prescriptive rules about what is acceptable in relations among individuals, groups, organizations, or other social units. Norms can be formalized in law, standards, or codes of behavior and backed up by institutionalized systems of rewards and sanctions. "Goals" are understood here as situation-specific descriptive states about the desirable future state of the H-CPS and its environment, which are chosen and/or prioritized to be conformant with the prevailing moral system.

[22] See e.g. recommendations of the National Academies of Sciences, Engineering, and Medicine in *Fostering Responsible Computing Research: Foundations and Practices*. Washington, DC: The National Academies Press, 2022, https://doi.org/10.17226/26507.

For technical systems, we assume that such moral beliefs are either entered into the system at design time or adopted automatically based on current location of the system. However, this strong assumption must be relativized: it is in general only possible to give a fuzzy or probabilistic interpretation either of moral beliefs or the more specific ethical principles of socially acceptable behaviors. In fact, even regulations such as the German Traffic Regulations or the law governing accountability decisions in traffic accidents do not allow for such an unambiguous model. Hence, we can only expect fuzzy moral beliefs to be represented in technical systems, such as learned by model extraction techniques.

Organizations are understood to be legal entities operating within the regulations and laws applicable to their application domain, and often these organizations have internal vision/mission statements as well as soft principles of operation, which then jointly define their moral system.

Humans interacting with safety-critical or industry-critical technical systems are exposed during their training to the relevant regulations and operational procedures governing the interaction with the physical system.

The control component at the moral system layer reflects the fact that norms and values can be temporarily and/or partially disregarded based on the current set of descriptive beliefs of lower levels or on the current goals for the situation. The control component thus allows the ego to adopt goals based on either broadly shared ethical codes or rules or on lower-level goals.

Examples

- An extreme stress situation of a driver might cause the driver to disregard traffic rules and adapt an extremely aggressive driving style, thus temporarily overriding the otherwise dominant influence of his/her moral system. To this end, the meta-model allows feedback from any of the lower levels to the control component of the moral system layer.
- Detecting a deer during a night drive in a situation of oncoming traffic may propagate the stress resulting from this near crash situation from the reflex layer to the moral system layer, causing a lane-change maneuver to avoid the deer, in spite of the fact that it violates traffic rules by forcing emergency braking of the opposing traffic.
- Receiving a call that her child is hospitalized in emergency conditions causes the mother to change strategy and drive to the hospital disregarding traffic regulations.
- A violation of integrity caused by a cyber-attack of the ego system might cause abnormal reactions that do not reflect its moral system.
- Being temporarily charged with rescue operations will allow an ego system to violate traffic regulations.

EGO 2 Reflection Layer

A key challenge in HCPS design is assuring a sufficiently high consistency of descriptive beliefs of cooperating systems, so as to allow them to realize the goals of the HCPS. Note that an ego system by itself, deprived of perception and communication, is not able to measure the confidence level of its beliefs. Mechanisms used to increase confidence are: 1) performing experiments validating descriptive beliefs, 2) using multiple sensor systems and sensor fusion and quantifying the degree of imprecision along such sensor chains, 3) sharing measured data with associated confidence levels across secure communication channels, and 4) failure detection and recovery. These mechanisms form necessary but not necessarily sufficient measures to estimate conservatively the currently achieved level of confidence of its descriptive beliefs, relative to environment assumptions, including intruder models and failure hypothesis. The ego system thus constantly reflects the plausibility of its descriptive beliefs by employing multiple forms of redundancy such as applicable physical laws, sensor fusion, and time-series analysis to either strengthen or weaken beliefs. Thus, the ego system constantly updates its current set of beliefs of possible worlds, and in particular it discharges some beliefs as unrealistic while at the same time creating new possible worlds. Control components of the ego system will consider all possible worlds. Control decisions will be evaluated with respect to consistency with the agent's moral system. This approach can be refined by assuming the availability of a metric of an omniscient observer, which we call precision, for measuring the discrepancy between reality and what the ego system believes to be true about the environment. The construction of such a measure is application dependent, but it follows principles in so-called metric temporal logics (see e.g. [1], [50]) which, intuitively, measure the degree of falsification of a formula, with *1* indicating complete satisfaction.

3.4 Lower Layers of the ego-system: from high-level planning to health state management

EGO 3 High-Level Planning

Each system follows a set of currently active high-level goals. These goals are dependent on the aggregation level of the system. For higher aggregation levels, these will subsume multi-optimization and trade-offs between subgoals (such as reduction of energy consumption, reduction of emissions, ensuring mobility, ensuring health care for all) down to reaching a particular destination at a certain time.

High-level planning determines, based on the current set of descriptive beliefs, if such goals are obtainable with the currently available resources. Control actions on this level include buy-in of additional resources (in our model, this buy-in would be leading to extensions of the capabilities of the ego system), reorganizing perception systems (such as obtaining more accurate beliefs or being able to observe additional artefacts), proactively maintaining the system, and reorganizing communication channels with other systems. Control actions on this level lead to a determination of a set

of prioritized goals to be implemented by lower levels. Note that goals may be subject to short-hand cancellations and revisions based on feedback from lower levels. Descriptive beliefs, goals and control actions are represented using high-level abstractions that are refined on lower levels.

Example

- Consider an emergency rescue scenario. A police officer has the role of evacuating a particular street and has as capability a police car equipped with a megaphone that makes it possible to announce the need for all pedestrians and vehicles to follow evacuation orders. When detecting a seriously injured person, the officer might contact the police headquarters to ask for authority to switch roles to perform emergency rescue operations using the emergency health treatment equipment in the officer's police car. If the officer is unable to guarantee sufficient first-level rescue treatments, he or she may ask for an emergency rescue helicopter to fly in, and in the meantime, switch to a role of guarding the injured person while guiding the helicopter pilot to a close landing approach. The police officer will pass the precise coordinates of the injured person to the rescue helicopter. The particular rescue helicopter deployed will be selected after taking into account its distance from the injured person, and, depending on the accessibility of the location, possibly its maneuvering capabilities determined by various physical parameters.

EGO 4 Strategy formation

A strategy for reaching a given goal determines the set of actions involved by the system for each change in the ego´s beliefs until the system reaches the state specified by the goal. The strategy is possibly extended by a justification of why the chosen actions are capable of guiding the ego system towards the desired goal.

The strategy level determines a strategy to achieve its goals by using the ego system's state as well as descriptive beliefs about its own health and integrity, its own capabilities, the state and capabilities of the environment, and the potential relationships between the ego and the environment. Strategies can be adversarial or cooperative with the behavior of other entities in the environment and can be formulated using optimization or reinforcement learning. See the companion paper [17] for a detailed discussion of strategies.

Technical systems with enough resources will consider all possible worlds in assessing the relative likelihood that alternate strategies or alternative tactics within an existing strategy can better achieve the system's goals. A strategy of a constituent system decides at any point in time the choice of action of its associated systems based on its descriptive and moral beliefs. The strategy assesses for a bounded time horizon the planned moves of the ego system for the effects of possible actions on the ego system and its environment. Humans will typically only explore possible evolutions of the more plausible beliefs, and they will use heuristics rather than a full exploration of the state of the entire HCPS system to decide on what approach to take to best achieve the current goals.

The following conditions can cause a cancellation of the strategy: unpredicted events (such as the detection of icy road conditions), asynchronous changes in the HCPS's state (such as from cyber-attacks or system failures as detected by the lowest level), and inconsistencies in beliefs as detected by the reflection layer,.

Example

If the autonomously driving car believes than an injured person is lying on its lane 50 meters ahead, and that there is either a truck or a tractor approaching on the opposite lane at a distance of 250 meter on a country road, it will have to assume the worst-case typical dynamics associated with trucks and tractors on country roads. Doing so will enable the system to assess whether it is safer to change the lane to avoid hitting the injured pedestrian or whether it is better not to change lanes.

For the purposes of this section,[23] a winning strategy in time horizon Δ is defined as achieving all its current safety goals and all its time-bounded reachability properties expiring in Δ within the next Δ seconds. The time horizon of a strategy will be typically chosen taking into account the assumed environment model and the short-term goals. To determine such a strategy, a system will use its descriptive beliefs about the environment model to assess (approximately) the future evolution of the real-world state based on its currently believed state up to the time horizon of the strategy.

If this analysis shows that no winning strategy exists, then the goals that are violated are flagged as unachievable, and the ego system S might choose to send cooperation requests to neighboring systems. In doing so, the goals of S can only be achieved by their cooperation—that is, by restricting their behavior to the activation of capabilities beneficial to S. Formally, if another system accepts a cooperation request, it adds the current goals of the system requesting help to its own list of goals, thus adapting its own current strategy to take into account these new goals. See [17] Part III of this publication series for a game-theoretic formal definition of these concepts.

Typically, cooperating systems will exchange their respective descriptive beliefs about the real world and agree to a shared view using belief fusion (a generalization of sensor fusion). Intuitively, belief fusion resolves inconsistent beliefs based on confidence levels, and it simply extends the existing beliefs of one system with beliefs about objects not previously observed by the other system. This coordination includes in particular exchange of descriptive beliefs about

23 For formal definitions, see the companion paper [17] on a game theoretic semantics of RA(HCPS)

the prevailing environment dynamics. Strategy synthesis in cooperating system is thus carried out based on consistent beliefs about the environment. Cooperating systems will temporarily include a subset of goals of the ego system in their own goals, until the duration window of the coalition has expired. An actual implementation would also consider the trustworthiness of a system, and it would only ask trustworthy systems for cooperation.

Assessing and redefining system goals in light of new information can also include redefining the target of action.

Examples:

- The driver (as ego system) asks its car to compensate for the driver's impaired vision on the left side, which causes a condition half-blindness.
- The autonomous car (as ego systems) asks the driver to double check for a person on the pavement because the combination of extreme rain and night time lighting is impairing the car's own perception. Such cooperation request rests on beliefs of the environment's capabilities, such as "knowing," as part of the ego´s environment belief, that the driver is fully alert and already scanning this area for potential risks, as would be deducible from indoor sensors observing the driver state.

In assessing the realizability of goals, we assume that systems in the environment act rationally, i.e., that they will not consciously obstruct the ego-system in following its goals, unless it is necessary to achieve their own goals (see [17] Part III for a formal definition). This condition holds for all systems that are not classified as adversarial systems. Descriptive beliefs that the environment is adversarial are created both internally at the reflective layer (through anomaly detection) as well as through communication with other CPSs in the environment. Classifying an environment as adversarial automatically changes its beliefs about the anticipated dynamics and capabilities of such systems in a way reflecting the expected aggressive behavior, such as based on previously observed attacker models. A system is trustworthy if its actions and the outcomes of the actions are continuously consistent with higher-level goals as given by the moral layer and the reflection layer.

Note that a strategy believed to be winning might not be winning in reality: the synthesis of the strategy is by necessity based on the system´s beliefs about the environment and itself. If its beliefs are of poor quality, then following the strategy will lead to situations where the actually observed state at some point in time will differ from the state the strategy expected to reach at that point in time. In such situations, in which the ego-system´s prediction about the future system state turns out to be incorrect, the execution of the strategy must be abandoned, and a new strategy must be synthesized based on the updated beliefs. This learning step will typically also involve updating beliefs about parameters of the environment mode. This change is accomplished by comparing sequences of actually observed sensor data to the expected beliefs, based on the internal representation of the environment dynamics in the current environment mode. Updating may also include learning about mode-switches in the environment model when parameter-fitting methods are not able to explain the deviations between expected and actually observed trajectories.

EGO 5 Maneuver Selection

This level selects for each point in time when a maneuver or action will be executed, based on the following factors: the current strategy, the current state of the HCPS and its capabilities, the current set of beliefs about the state and actions of the environment (such as in the case of environmental risks that induce a lane change), and the response to critical deviations of the environment from its anticipated dynamics. Note that predictions of such dynamics are of statistical nature only. Coalition formation may agree to exchange intentions about next planned steps, according to the prevailing strategy, and doing so may reduce the number of cancellations of maneuvers based on possible but unlikely actions. To this end, descriptive beliefs include beliefs regarding intentions of neighboring systems. Such beliefs may also be updated based on perceived cues. Maneuvers may be also be cancelled by the health-state and integrity monitor, leading to fail-safe or fail operational maneuvers, or, if not available, to minimal risk maneuvers.

Examples

- the observation that a child who has lost control over a ball that has rolled into the street will induce the belief that the child's intent will change from staying on the pavement to catching the ball.

EGO 6 Reflexes – Low Level control

For technical ego-systems, this level is the well understood level of control design for performing the chosen maneuvers. It relies on its own state and beliefs of ego´s health state and capabilities, as well as descriptive beliefs about the state of surrounding systems of the environment, including environmental conditions, and their dynamics. Note that such beliefs subsume as special case the state of the operator or the state of the driver, and elaborate mechanisms can be employed to infer a wide range of driver states, such as the driver's current level of frustration, trust, fatigue, overload, and alertness. These inferences allow the HCPS to adapt dynamics for low-level control.

For human ego systems, the reflex layer subsumes all activities that are carried out purely subconsciously. The activities range from reflexes such as panic reflexes to automatic direction of attention to fast moving objects in side view to emotion formation, reflex behavior, formation of long-term memory, and other functions anchored in the limbal system. We can describe the effect of training of professionals in the reference architecture as follows: activities that initially are carried out on the cognitive maneuver level can after sufficient exposure carried out subconsciously, thus freeing the cognitive level to perform other tasks.

Examples

- High-end cars feature emergency braking systems that pick up critical environmental dynamics and override all higher-level activities, focusing on the execution of a minimal risk maneuver.
- Skilled drivers have internalized control for maintaining the vehicle in the lane and keeping a safe distance from other vehicles on highways.

EGO 7 Integrity- and Health-State-Management

For technical systems, this component maintains the integrity of all data relevant for decision making on all hierarchy levels, using the range of techniques of security research for intrusion detection and protection against cyber-attacks. The component also ensures the safety and availability of all key functions of the ego systems against all failures specified in the fault hypothesis. To do so, it uses the range of techniques from safety research, including implementation of never-give-up strategies to handle types of attacks and failures not previously encountered.

Humans learn to compensate up to a certain level of sensory-motor disturbance through a combination of cognitive processes that complement a lack of sensory information. They do so by learning to use other sensors and/or training, possibly supported by assistive technology.

3.5 Layer Structure of the environment of the ego system

Because the environment of the ego system is comprised of systems that have this same hierarchical structure for their internal organization, *ego(ENV)* is canonically structured corresponding to the layers of the ego system. ENV3- ENV6 thus provide structure in the space of beliefs of the ego system:

ENV3 Environment Plans

Subsumes all knowledge about plans of the environment relevant for the planning of the ego system.

Examples

- If the environment is an emergency rescue vehicle, its goal is to reach the place of accident in the shortest possible time frame. Knowing that this vehicle is an emergency rescue vehicle is key to the ego system – it follows different traffic rules and different dynamic models and imposes mandatory cooperation on all vehicles in its trajectory
- Today's mechanisms of creating awareness about emergency vehicles are expected to be extended, such as by communicating the planned routes to surrounding vehicles, and/or to clear street segments ahead from potentially blocking vehicles.
- As a second example, city-wide crisis management systems rely on knowledge about which resources, held by other stakeholders at various locations in the city, will be deployed to address the crisis, so as to coordinate plans and resource allocations. Because an ego system for crisis management will be part of a virtual overarching system for crisis management, it will have access rights to obtain such information from other subsystems of the crisis management system.

ENV4 Environment Strategies

Subsumes all knowledge about all environment systems relevant for strategy synthesis of the ego system, such as their individual goals and strategies to achieve these goals and their relationship to the ego system that may be neutral, cooperative, or adversarial. Sharing such information is dependent on being part of a common superstructure, entering temporary cooperation agreements, or both.

ENV5 Environment Capabilities and Dynamics

Having knowledge about the environment's capabilities and dynamics is a fundamental prerequisite for maneuver selection in any application domain. Such models are estimated based on validated statistical models of dynamics of systems in a given operational context and a given location. The models are available at design time and monitored during operation time, possibly using learning-enabled components to update such models at run-time.

Examples

- the estimated behavior of cars at the Arc de Triumph will not assume any respect of lanes, and aggressive strategies for entering the desired lane, differing radically from the expected behavior of cars in a roundabout in Germany.
- Cars on most highways in US are expected to drive more slowly than 70 mph, and cars are passing both on the left as well as the right side. In contrast, cars on highways in Germany are expected to overtake slower cars only on the left, and large segments of highways have no speed-limit, leading to highly variant dynamics of environment traffic.

ENV6 Environment State and Class

Identification of the class of surrounding systems according to an application-dependent standardized ontology is a mandatory requirement for determining safe maneuvers. Such ontologies are currently developed as part of the VDA Leitinitiative (see https://www.vda.de/en) for autonomous driving in Germany. Estimating the state of other systems includes in particular estimating descriptive beliefs that this system has about the ego system´s state and beliefs.

Example

- The ego car will choose a particular trajectory when approaching a pedestrian crossing "indicating" to waiting pedestrians that the ego vehicle has identified their intent to cross the street.

ENV7 Uncontrolled EGO State, Capabilities and Dynamics

The ground truth about part of the ego state, its capabilities, and its dynamics are all part of *ego(ENV).* This may be surprising on first reading because we interpret the concept of "environment" as being any uncontrolled entity that is potentially acting adversarial to the ego system. As is well known, the very physical nature of an ego system, be it a CPS or a human, makes the ego system vulnerable to uncontrollable failures. Hence, the physical aspect of the ego system has to be considered part of the environment. Because such failures can impact any of the ego´s subsystems, they may lead to false perceptions of its state, false perception of its capabilities, and false perceptions of its dynamics. Thus, a central part of the ego system is to monitor its health and integrity state, as done by EGO7.

4. CONCLUSION

As the degree of automation in CPSs is rising, designing HCPSs requires a new level of understanding that addresses the many facets in human-CPS interactions. This holds true in particular for the class of safety-critical or industry-critical systems addressed in this paper, because misunderstanding and lack of cooperation can cause disastrous effects. It calls for design processes, which must go well beyond the current state of industrial practice. The design must ensure that the key messages relevant for cooperation and controlling potentially chaotic situations are actually perceived by the human operator(s) in time, and the CPS understands the operator´s strategies, intentions, and assessments of the process to be controlled jointly. Often, such challenges can only be handled by highly interdisciplinary teams. Complexities of human-system interactions call for end-to-end verification, such as verifying that critical situations are not only identified and passed through the human computer interface, but also that they actually lead to percepts that allow cognitive analysis by the human operator.

This study is intended to serve as a blueprint for designing such systems. The structure of the reference architecture in multiple layers directs the attention of designers to consider which levels are involved in their problems. The approach also generates questions such as whether joint actions are compliant with prevailing societal values; compliant with existing regulatory frameworks; and consistent with respect to goals, plans, and strategies among the involved actors. It highlights the need to assess the extent to which mutual beliefs (of beliefs of beliefs ….) about the other´s goals, plans, strategies, and perceptions of the environment are sufficiently consistent and sufficiently precise. Throughout the study, we have given numerous examples from different application domains to highlight the relevance of such types of assessments.

5. ACKNOWLEDGEMENTS

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.

6. REFERENCES

[1] Houssam Abbas, Georgios Fainekos, Sriram Sankaranarayanan, Franjo Ivančić: “Probabilistic Temporal Logic Falsification of Cyber-Physical Systems,” *ACM Transactions on Embedded Computing Systems*, Volume 12,Issue 2, 95pp 1–30, May 2013 https://doi.org/10.1145/2465787.2465797

[2] Erzana Berani Abdelwahab, Martin Fränzle. 2021. A Sampling-Based Approach for Handling Delays in Continuous and Hybrid Systems. In it - Information Technology, vol. 63, iss. 5-6, pages 289-298.

[3] Special Issue on the Synthesis of Cyber-Physical Systems, ACM Transaction on Embedded Computing Systems, 2012.

[4] Janna Anderson, Lee Rainie. 2018. Artificial intelligence and the future of humans. Pew Research Centre, 2018. https://www.pewinternet.org/2018/12/10/artificial-intelligence-and-the-future-of- humans/. Accessed 25 Sept 2019.

[5] Peter M. Asaro. 2019. AI ethics in predictive policing: From models of threat to an ethics of care. IEEE Technology and Society Magazine, 38(2), June 2019, pages 40–53. https://doi.org/10.1109/MTS.2019.2915154

[6] Leonie Beining, Peter Bihr, Stefan Heumann. 2020. Towards a European AI & Society Ecosystem - Why we need it and how to empower it to shape Europe's way on AI. In Stiftung neue Verantwortung, 2020, Berlin. https://www.stiftung- nv.de/sites/default/files/towards_a_european_ai_society_ecosystem_0.pdf)

[7] Klaus Bengler, Werner Damm, Andreas Luedtke, Jochem Rieger, et al. 2023. A reference architecture for human cyber physical systems- part II: Fundamental design principles for human-cps interaction in Transactions on Cyber-Physical Systems X(Y), ACM (Association for Computing Machinery), New York, NY, United States.

[8] Bundesamt für Sicherheit in der Informationstechnik (BSI) (Ed.), AI Cloud Service Compliance Criteria Catalogue (AIC4), https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud- Service-Compliance-Criteria-Catalogue_AIC4.html

[9] Bundesministerium des Innern, für Bau und Heimat (Ed.), Gutachten der Datenethikkommission der Bundesregierung, Berlin, 2019.

[10] Bundesregierung Deutschland (Ed.), Strategie Künstliche Intelligenz der Bundesregierung, 2020, https://www.bmbf.de/files/Nationale_KI-Strategie.pdf

[11] Bundesverband Digitale Wirtschaft (BVDW) e.V. (Ed.), Acht Leitlinien für künstliche Intelligenz - Leitlinien des BVDW, Berlin, 2019.

[12] Mingshuai Chen, Martin Fränzle, Yangjia Li, Peter Nazier Mosaad, Naijun Zhan. 2021. Indecision and delays are the parents of failure - taming them algorithmically by synthesizing delay-resilient control. In Acta Informatica 58(5), pages 497-528.

[13] Nerdia Creswick, Johanna Irene Westbrook. 2007. Social network analysis of medication advice-seeking interactions among staff in an Australian hospital. In int J Med Inform., ITHC 2007 special issue, doi:10.1016/j.ijmedinf.2008.08.005

[14] Peter Dabrock, Michael Decker, Werner Damm, Armin Grunwald, Jessica Heesen, Klaus Heine, Detlef Houdeau, Tobias Matzner, Jörn Müller-Quade, Peter Rost, Thomas Schauf, Katharina Zweig, Stefan Wrobel. 2021. Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten. White Paper of the German Platform on Learning Enabled Systems, August 2021, to appear, 31 pages.

[15] Werner Damm, Martin Fränzle, Willem Hagemann, Astrid Rakow, and Mani Swaminathan. 2020. Assuring Confidence in the Perception Chain of Highly Automated Vehicles. Technical report, Mai 2020, Universitaet Oldenburg.

[16] Werner Damm, Martin Fränzle, Andreas Lüdtke, Jochem W. Rieger, Alexander Trende, Anirudh Unni. 2019. Integrating Neurophysiological Sensors and Driver Models for Safe and Performant Automated Vehicle Control. In Mixed Traffic, IV 2019, pages 82-89.

[17] Werner Damm, Martin Fränzle, Alyssa J. Kerscher, Forrest Laine, Forrest, et al. 2023. A reference architecture for human cyber physical systems- part III: Semantic foundations, in Transactions on Cyber-Physical Systems X(Y), ACM (Association for Computing Machinery), New York, NY, United States.

[18] Werner Damm, Alberto L. Sangiovanni-Vincentelli. 2015. A conceptual model of system of systems. Proceedings Second International Workshop on the Swarm at the Edge of the Cloud, CPS Week 2015, pages 19-27.

[19] Werner Damm, Johannes Helbig, Peter Liggesmeyer, Philipp Slusallek. 2021. Trusted AI: Why We Need a New Major Research and Innovation Initiative for AI in Germany and Europe. White paper submitted to German Federal Ministry of Education and Research, March 2021, 41 pages.

[20] Werner Damm and Peter Heidl. 2021. SafeTRANS Roadmap on Safety, Security, and Certifiability of Future Man-Machine Systems. SafeTRANS, Jan 2021, 64 pages, available from www.safetrans-de.org

[21] Lida David, Jan Maarten Schraagen. 2018. Analyzing communication dynamics at the transaction level: the case of Air France Flight 447. In Cognition, Technology & Work. 20. 10.1007/s10111-018-0506-y

[22] Frederik Diederichs, Arun Muthumani, Alexander Feierle, Melanie Galle, Lesley-Ann Mathis, Valeria Bopp-Bertenbreiter, Harald Widlroither, Klaus Bengler. 2022. Improving Driver Performance and Experience in Assisted and Automated Driving With Visual Cues in the Steering Wheel. In IEEE Trans. Intell. Transp. Syst. 23(5), pages 4843-4852.

[23] European Commission, High-Level Expert Group on Artificial Intelligence (Ed.), Ethics Guidelines for Trustworthy AI, Brussels, 2019.

[24] European Commission (Ed.): White Paper on Artificial Intelligence - A European approach to excellence and trust, Brussels, 2020. https://ec.europa.eu/info/sites/info/files/commission- white-paper-artificial-intelligence-feb2020_en.pdf

[25] European Union Aviation Safety Agency (EASA) (Ed.), Artificial Intelligence Roadmap - A human-centric approach to AI in aviation, Köln, 2020.

[26] Shenghua Feng, Mingshuai Chen, Naijun Zhan, Martin Fränzle, Bai Xue, 2019. Taming Delays in Dynamical Systems - Unbounded Verification of Delay Differential Equations. In Computer Aided Verification - 31st International Conference, CAV 2019, July 15-18, 2019, New York City, NY, USA, pages 650-669.

[27] Bernd Finkbeiner, Leander Tentrup. 2015. Detecting Unrealizability of Distributed Fault-tolerant Systems. In Log. Methods Comput. Sci. 11(3).

[28] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Christopher Nagy, Madhulika Srikumar. 2020. Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. The Berkman Klein Center for Internet & Society Research Publication Series, Research Publication No. 2020-1, Jan 2020.

[29] James Fordyce, Fidela S j Blank, Penelope Pekow, Howard A. Smithline, Georg Ritter, Stephen Gehlbach, Evan Benjamin, Philipp L. Henneman. 2003. Errors in a busy emergency department. In Ann Emerg Med. 2003, 42 (3), pages 324-333. 10.1016/S0196-0644(03)00398-6.

[30] Yolanda Gil and Bart Selman. 2019. A 20-Year Community Roadmap for Artificial Intelligence Research in the US. Computing Community Consortium (CCC) and Association for the Advancement of Artificial Intelligence (AAAI), August 2019. arXiv:1908.02624 https://cra.org/ccc/ resources/workshop-reports/

[31] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. In Minds & Machines 30, pages 99– 120, (https://link.springer.com/content/pdf/10.1007/s11023-020-09517-8.pdf)

[32] John C. Harsanyi. 1967. Games with incomplete information played by “Bayesian” players, I–III Part I. The basic model. In Management science 14.3, pages 159-182.

[33] David Hess. 2020. The Sociology of Ignorance and Post-Truth Politics. In Sociological Forum 35(1), 241-249, 2020, http://dx.doi.org/10.1111/socf.12577.

[34] David J. Hess, Dasom Lee, Bianca Biebl, Martin Fränzle, Sebastian Lehnhoff, Himanshu Neema, Jürgen Niehaus, Alexander Pretschner, and Janos Sztipanovits. 2021. A Sociotechnical Design Perspective on Responsible Innovation: Perspectives on Problem Finding for Multidisciplinary Research on Digitized Energy and Automated Vehicles. Journal of Responsible Innovation 8(3), pages 421-444.

[35] David J. Hess. 2022. Undone Science and Social Movements: A Review and Typology. In Matthias Gross and Linsey McGoey (ed.), The Routledge International Handbook of Ignorance Studies. Routledge. Second edition.

[36] David J. Hess. 2022. Undone Science and Smart Cities: Civil Society Perspectives on Risk and Emerging Technologies. Johannes Glückler, Heinz-Dieter Meyer, Laura Suarsana (eds) Knowledge and Civil Society (Knowledge and Space, Vol 17). In Cham: Springer International, pages 57-73.

[37] Heads of Medicines Agencies, European Medicines Agency (Ed.), HMA-EMA Joint Big Data Taskforce Phase II report: Evolving Data-Driven Regulation, 2019.

[38] Sebastian Houben et al. 2020. Inspect, Understand, Overcome: A Survey of Practical Methods for AI Safety. In KI Absicherung / https://www.ki-absicherung-projekt.de/fileadmin/KI_Absicherung/Downloads/KI-A_20201221_Houben_et_al_-_Inspect__Understand__Overcome.pdf

[39] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, First Edition, IEEE, 2019.

[40] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, https://standards.ieee.org/industry-connections/ec/autonomous-systems/

[41] IEEE, Symbiotic Autonomous Systems - White Paper III, IEEE, 2019.

[42] ISO/IEC AWI TR 5469, Artificial intelligence — Functional safety and AI systems; https://www.iso.org/standard/81283.html

[43] Anna Jobin, Marcello Ienca, Effy Vayena. 2019. The global landscape of AI ethics guidelines. Nat Mach Intell 1, pages 389–399, https://doi.org/10.1038/s42256-019-0088-2

[44] Tobias Krafft, Marc Hauer, Lajla Fetic, Andreas Kaminski, Michael Puntschuh, Philipp Otto, Christiph Hubig, Torsten Fleischer, Paul Grünke, Rafaela Hillerbrand, Carla Hustedt, Sebastian Hallensleben. 2020. From Principles to Practice - An interdisciplinary framework to operationalise AI ethics. AI Ethics Impact Group led byVDE / Bertelsmann Stiftung.

[45] Dasom Lee and David J. Hess. 2022. Public Concerns and Connected and Automated Vehicles: Safety, Privacy, and Security. In Humanities and Social Sciences Communications (Springer Nature), special issue on CAVs and society, https://doi.org/10.1057/s41599-022-01110-x.

[46] Mathias Lechner, Ramin Hasani, Alexander Amini, Thomas A. Henzinger, Daniela Rus, Radu Grosu. 2020. Neural circuit policies enabling auditable autonomy. In Nature Machine Intelligence, Volume 2, Pages 642–652, October, 2020.

[47] Andreas Lüdtke. 2015. Wege aus der Ironie in Richtung ernsthafter Automatisierung. Botthof, A., Hartmann, E. (eds) Zukunft der Arbeit in Industrie 4.0. In Springer Vieweg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-45915-7_13.

[48] National Security Commission on Artificial Intelligence. Final Report, 2021. https://reports.nscai.gov/final-report/table-of-contents.

[49] Julian Nida-Rümelin, Nathalie Weidenfeld. 2018. Digitaler Humanismus: Eine Ethik für das Zeitalter der Künstlichen Intelligenz. In Piper; 4. Edition, September 04, 2018, München, Germany.

[50] Jo¨el Ouaknine, James Worrell. 2005. On the Decidability of Metric Temporal Logic, Logic in computer science (LICS 2005); proceedings. 20th Annual IEEE Symposium on Logic in Computer Science (LICS' 05), Chicago, IL, Computer Society Press 2005.

[51] Plattform Lernende Systeme (Ed.), Zertifizierung von KI Systemen – Positionspapier, 2020.

[52] Plattform Lernende Systeme (Ed.), Kritikalitätsbewerung von KI Systemen – Positionspapier, 2020.

[53] 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), pages 257–266.

[54] Ian B. Rhodes, David G. Luenberger. 1969. Differential games with imperfect state information. IEEE Transactions on Automatic Control 14.1 (1969), pages 29-38.

[55] Matthias Rolf, Nigel Crook, Jochen Steil. 2018. From social interaction to ethical AI: a developmental roadmap. 2018 Joint IEEE 8th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob), Tokyo, Japan, pages 204-211, doi: 10.1109/DEVLRN.2018.8761023, 2018.

[56] Mark Ryan. 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. Sci Eng Ethics, 26, pages, 2749– 2767, 2020, https://link.springer.com/content/pdf/10.1007/s11948-020-00228-y.pdf

[57] SafeTRANS (Ed.), Safety, Security, and Certifiability of Future Man-Machine Systems – Roadmap, Oldenburg, 2019.

[58] Shankar Sastry, Jonas Sztipanovits, Ruzena Bajcsy, Helen Gill. 2003. Model-Based Design of Embedded Systems: Scanning the Issue, Proceedings of the IEEE, Vol. 91, No.1., pp. 4-10, January, 2003.

[59] Special Issue on Cyber Physical Systems, Proceedings of the IEEE, Vol 100, January 2012.

[60] Special issue on Design Automation for Cyber-Physical Systems. Proceedings of the IEEE, Vol 106 September 2018.

[61] Sulayman K Sowe, Martin Fränzle, Jan-Patrick Osterloh, Alexander Trende. 2019. Challenges for Integrating Humans into Vehicular Cyber-Physical Systems. In 17th edition of the International Conference on Software Engineering and Formal Methods, September 16 – 20, 2019, Oslo, Norway.

[62] Janos Sztipanovits, Ted Bapty, Ethan Jackson, Xenofon Koutsoukos, Zsolt Lattman, Sandeep Neema. 2018. Model and Tool Integration Platform for Cyber-Physical System Design, Proceedings of the IEEE, special issue on Design Automation for Cyber-Physical Systems, 2018.

[63] Janos Sztipanovits, Xenofon Koutsoukos, Gabor Karsai, Shankar Sastry, ClaireTomlin, Werner Damm, Martin Fraenzle, Jochem Rieger, Alexander Pretschner, Frank Koester. 2019. Science of Design for Societal-Scale Cyber-Physical Systems: Challenges and Opportunities, Journal on Cyber-Physical Systems, Volume 5, 2019 – Issue 3. Taylor& Francis. DOI: 10.1080/23335777.2019.1624619

[64] Matthew U. Scherer. 2016. Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. Harvard Journal of Law & Technology, Vol. 29, No. 2, Spring 2016.

[65] Manuel Schrepfer. 2013. Ich weiß, was du meinst! In Theory of Mind, Sprache und kognitive Entwicklung. AVM Verlag, München 2013, ISBN 978-3-86924-502-7.

[66] Scott Thiebes, Sebastian Lins, Ali Sunyaev. 2021. Trustworthy artificial intelligence. Electronic Markets, pages 447-464, https://link.springer.com/content/pdf/10.1007/s12525-020-00441-4.pdf, 2020.

[67] Alexander Trende, Anirudh Unni, Lars Weber, Jochem W. Rieger, Andreas Lüdtke. 2019. An investigation into human-autonomous vs. human-human vehicle interaction in time-critical situations. Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, pages 303-304.

[68] Alexander Trende, Franziska Hartwich, Cornelia Schmidt, Martin Fränzle. 2020. Improving the Detection of User Uncertainty in Automated Overtaking Maneuvers by Combining Contextual, Physiological and Individualized User Data. In HCI (40) 2020, pages 390-397.

[69] Alexander Trende, Ina Krefting, Anirudh Unni, Jochem W. Rieger, Martin Fränzle. 2022. A Case-Study for a Human-Centered Approach to Traffic Management Systems. HCI- 24th International Conference on Human-Computer Interaction (40) 2022, pages 259-266.

[70] UNESCO (Ed.), First Draft Recommendation on the Ethics of AI, September 2020.

[71] VDE-AR-E 2842-61, Development and trustworthiness of autonomous/cognitive systems, https://www.vde-verlag.de/standards/1800575/e-vde-ar-e-2842-61-2-anwendungsregel- 2020-07.html

[72] Wolfgang Wahlster, Christoph Winterhalter (Ed.), Deutsche Normungsroadmap Künstliche Intelligenz, https://www.din.de/-resource/blob/772438/6b5ac6680543eff9fe372603514be3e6/normungsroadmap-ki-data.pdf

[73] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, Stephen Cave. 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. Nuffield Foundation 2019, London. https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

[74] Jeanett M. Wing. 2021. Trustworthy AI. Communications of the ACM, October 2021, Vol. 64 No. 10, pages 64-71.

ANNEX I:

Terminology for the Human Dimensions of the Reference Architecture

We suggest that the metamodel use the following terminology.

1. Prescriptive states

The term "prescriptive" is used to describe states in human-cyber-physical systems that refer to "ought" conditions or what states of a situation or scenario "should" be. Generally, we distinguish various subcategories:

Values are general principles that guide action. Values are about what is important, good, beneficial, or desirable. In some societies and some formal systems, values may be organized hierarchically.

"Norms" are prescriptive rules about what is acceptable in relations among individuals, groups, organizations, or other social units. Norms can be formalized in law, standards, or codes of behavior and backed up by institutionalized systems of rewards and sanctions. For example, if someone takes a life without cause, the person may be sanctioned as a murderer and punished. Norms can also be more informal, such as proper behavior in interpersonal relations or in public spaces. There can also be informal rewards or sanctions for social action that corresponds with acceptable social practices or customs.

Together, values and norms can be described as "morals." The term "Morality/moral" refers to a set of commonly accepted (evaluative or normative) standards (e.g., traditions, custom, conventions) that can vary between different places and historical times.

Moral beliefs: We take moral beliefs to be evaluative or normative judgments about what is desirable or morally right/wrong. Prescriptive beliefs can be characterized as valid, questionable, or invalid.

By contrast, "ethics/ethical" refers to the theoretical reflection and discussion of morality/moral values and norms with regard to their validity, acceptability and legitimacy. Sometimes the results of ethical reflections and discussions are expressed in ethical codes or systems. These can have varying scope from a broad code for general action to more specific codes. For example, professional associations often have codes of conduct. These codes or systems can acquire the status of soft laws so that people who violate them may be accused of unethical action or be faced with sanctions. As soft laws, ethical codes or systems can vary between different ethics bodies and national contexts.

Although ethical codes are widely used in professions, they may not provide enough detail to do more than provide general principles for use in projects like software design. Thus, one approach to bringing moral or ethical perspectives into software design focuses more on the design process and ensuring the participation of multiple stakeholders (including persons who represent accepted values for the projects such as safety, privacy, security, equity, and sustainability). Another approach is the one pursued in "machine ethics": The implementation of capacities for moral judgment/behavior in technical systems.

2. Descriptive states

The parallel with the prescriptive states is descriptive states. These states describe actual situations about the self or the world, or they make predictions about how situations are likely to change given current information or planned actions. (Other terms are also used. Some use the term "positive" in contrast with "normative.")

When applied to specific situations, a cyber-physical system can be programmed to attempt to realize specified states of a situation that are deemed desirable from the perspective of a system of prescriptive states as defined above. We use the term "goals" to refer to these situation-dependent descriptive states that are programmed into a system or that human actors have in a specific situation. Another term for "goals" is "objectives."

Descriptive beliefs will be understood here as general representations of "is" conditions. Descriptive beliefs can include descriptions of the actual state of the world (or, at a more local level, of a situation or system) and predictions about how the state of the world will change or not change in response to trends in existing states or planned actions of the self or anticipated actions of others. Beliefs can be characterized as valid, questionable, or invalid.

Justified descriptive beliefs (or "facts") are beliefs that have been evaluated for validity based on a system of analysis or assurance, similar to how ethical systems formalize and systematize moral norms and values. The system of justification of beliefs does not guarantee that the beliefs are accurate descriptions of a state of the world, but it improves the likelihood that the beliefs are "true." The system of justification varies. For example, scientific facts require a system of peer-review and widespread acceptance among the expert community to be considered adequately justified. Legal facts require a different type of justification based on the rules of jurisprudence in a particular region or country. Local knowledge facts require consensual validation by people who know and understand a community.

We use the term "assessment" to refer to the programming of a cyber-physical system that provides a picture of descriptive statements. Assessments are beliefs about an existing system or situation and future states. They usually are embedded in a system of justification to improve their accuracy.