A Reference Architecture of Human Cyber-Physical Systems – Part III: Semantic Foundations

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

The design and analysis of multi-agent human cyber-physical systems in safety-critical or industry-critical domains calls for an adequate semantic foundation capable of exhaustively and rigorously describing all emer-

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

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

gent effects in the joint dynamic behavior of the agents that are relevant to their safety and well-behavior. We present such a semantic foundation. This framework extends beyond previous approaches by extending the agent-local dynamic state beyond state components under direct control of the agent and belief about other agents (as previously suggested for understanding cooperative as well as rational behavior) to agent-local evidence and belief about the overall cooperative, competitive, or coopetitive game structure. We argue that this extension is necessary for rigorously analyzing systems of human cyber-physical systems because humans are known to employ cognitive replacement models of system dynamics that are both non-stationary and potentially incongruent. These replacement models induce visible and potentially harmful effects on their joint emergent behavior and the interaction with cyber-physical system components.

CCS Concepts: • **Human-centered computing** → *Interaction paradigms*; **HCI theory, concepts and models**; **Interaction design theory, concepts and paradigms; • Theory of computation** → **Timed and hybrid models**; **Interactive computation**;

Additional Key Words and Phrases: Cyber-physical systems, human cyber-physical systems, reference architecture, hybrid discrete-continuous dynamics, game theory, formal semantics

ACM Reference format:
Werner Damm, Martin Fränzle, Alyssa J. Kerscher, Forrest Laine, Klaus Bengler, Bianca Biebl, Willem Hagemann, Moritz Held, David Hess, Klas Ihme, Severin Kacianka, Sebastian Lehnhoff, Andreas Luedtke, Alexander Pretschner, Astrid Rakow, Jochem Rieger, Daniel Sonntag, Janos Sztipanovits, Maike Schwammberger, Mark Schweda, Alexander Trende, Anirudh Unni, and Eric Veith. 2024. A Reference Architecture of Human Cyber-Physical Systems – Part III: Semantic Foundations. *ACM Trans. Cyber-Phys. Syst.* 8, 1, Article 4 (January 2024), 23 pages.
https://doi.org/10.1145/3622881

1 INTRODUCTION

Model-based methods, dating back to the 1970s [4], have become an established means of exhaustively assessing behavioral properties of complex engineered systems, covering both functional properties like system safety across all conceivable application scenarios and so-called non-functional properties like timing or energy consumption. These methods achieve their analytic power by being semantically well-founded, i.e., by providing a rigorous and comprehensive mathematical or computational representation of the possible behaviors of the system under investigation. Such model-based methods are well-established across pure engineering domains (such as the construction of digital circuitry or embedded software) and across the combination of these domains and interaction with environments. Furthermore, these established methods are amenable to rigorous mathematical descriptions, like in the control of physical processes featuring a closed-form mathematical model of system dynamics. However, human cyber-physical systems are just starting to come into the purview of such methods. To extend these approaches to human cyber-physical systems, the model-based behavioral analysis of these systems requires establishing pertinent cognitive models and their seamless integration with models of cyber-physical system dynamics. This approach to analysis not only integrates human-machine interfaces and cognitive modeling of human behavior but also includes their emergent joint behavior within the purview

Strasse 1, 26129 Oldenburg/Germany; e-mail: daniel.sonntag@uni-oldenburg.de; E. Veith, OFFIS e. V., Escherweg 2, 26121 Oldenburg/Germany; e-mail: eric.veith@offis.de.

2378-962X/2024/01-ART4
https://doi.org/10.1145/3622881

of formal analysis and synthesis techniques. Achieving this goal requires us to specify a correspondingly expressive reference architecture for safety-critical or industry-critical human cyber-physical systems and to solidly anchor the architecture in a firm and unambiguous semantic basis.

Part I of this set of articles [6] addressed the first of the two prerequisites of rigorous model-based methods for safety-critical or industry-critical human cyber-physical systems, namely the provisioning of an adequate reference architecture. In Part III, we complement Part I by providing the corresponding semantic foundation. As human cyber-physical systems comprise cyber-physical system components, which in turn are a superset of hybrid discrete-continuous systems, the semantic foundation must generalize the pertinent models of hybrid discrete-continuous dynamics as encountered in smart systems, and it must blend continuous control with discrete supervisory decisions. Although adequate mathematical models for such hybrid discrete-continuous behavior, in particular several variants of hybrid automata [1, 9, 14], have evolved over the last three decades, they are not in themselves sufficient for covering the behavioral dynamics of human cyber-physical systems. It can be expected that humans interacting with their cyber-physical environment exhibit signs of rational or bounded rational behavior beyond the scope of these models. Furthermore, previous research (e.g., [8, 10]) has shown that even engineered systems cannot adequately be modeled in pre-existing hybrid systems formalisms if they employ rational decision-making. Our semantic base model therefore must adopt the wisdom and modeling approaches of game theory, especially the theory of durational and interactive dynamic games played under rational strategies [2, 9, 14], concerning the strategy finding and the resultant emergent dynamics of rational interacting agents.

However, such game theoretical approaches are not sufficient to cover our domain in the sense of providing a comprehensive mathematical model capable that rigorously describes all relevant effects that impact the joint system dynamics, i.e., the joint emergent dynamics of all agents in a complex human cyber-physical system. The notable shortcoming of most if not all mathematically concise dynamic game models of interactive multi-agent behavior is that they tend to assume both persistence and joint knowledge of the underlying game structure, i.e., of the game arena [20] detailing available actions and their consequences. Uncertainties in the evolution of the game are frequently supported but tend to apply only to individual evaluation of the current game state (which may only be partially observable to the individual agents) or to the consequences of actions (which may not fully be determined, rendering the resulting game state ambiguous), yet not to the structure of the game itself. We argue that this approach is too limited: humans are known to employ cognitive replacement models of system dynamics [3, 15] that are non-stationary – in that they are developed and updated based on observing the interacting system – and potentially incongruent in that different humans may develop different models or may concurrently refer to different revisions of the same model. As the local game model believed to be true by an agent has a fundamental impact on the individual selection of a rational behavioral strategy and on the individual assessment and justification of the strategy's safety, conflicts and inconsistencies potentially inducing hazardous joint emergent behavior may remain unnoticed when applying semantic models not reflecting the local nature of game arenas.

We address this problem by suggesting a semantic foundation for multi-agent human cyber-physical systems where each agent has her own, potentially dynamic, copy of the game arena together with the possibility of performing updates on that copy. We thus render game arenas, or rather local beliefs on the game arena, first-class members of an agent's state set. It should be noted that this addition may induce the effect that, in contrast to agent-local beliefs on other state components, there no longer is necessarily a corresponding ground-truth concerning the game structure: the agent-local beliefs on the game arena approximate the actual joint game arena of all the agents, which in turn is induced by the options believed to be available and the consequential strategy selections by the various agents locally.

In this article, we outline a corresponding foundational semantic model by first demonstrating the necessity of its elements by means of sample safety analyses in Section 2, building on the cockpit interaction instantiation of the reference architecture of Section 2 of Part II [3] of this series of articles, and then describing the elements and their interactions more formally in Section 3. The resulting model renders various forms of belief and of collected evidence into local state components of individual agents, namely belief and evidence about unobservable parts of physical state, belief and evidence about the state and plan of other agents, and belief and evidence about the overall game structure and the behavioral options available within it. It is our firm belief that the resulting integration of beliefs and evidence, especially about other agents and their strategic behavior plus its justification, is a prerequisite for obtaining sound verdicts about the emergent joint behavior of ensembles of agents. As it stands, the model is still abstract and devoid of any tooling. We communicate it as a basis for discussion that can contribute to the development of the state of the art in modeling of safety-critical or industry-critical multi-agent human cyber-physical systems.

2 A SAMPLE APPLICATION OF ANALYSIS METHODS: HAZARD ANALYSIS OF THE INSTANTIATION OF THE REFERENCE ARCHITECTURE TO COCKPIT COMMUNICATION

In this section, we show how the expressiveness of the reference architecture (see Part I [6]) in anchoring beliefs at all layers allows the capture of safety critical inconsistencies in beliefs. In terms of the semantic foundation of RA(HCPS), these inconsistencies in belief will induce inconsistencies between the assumed states of other players and also between the local arenas employed by players within a team for locally justifying their behavioral decisions. Such inconsistencies strongly impair the cooperation in what is called “groups of systems” in Section 2.1 of Part I [6], as discussed in Section 3.1 in the companion paper [3], where multiple human cyber-physical systems team up to jointly control one process.

We exemplify this for the group of systems formed by pilot, first officer, second officer, autopilot, and radio control in the cockpit of an aircraft, building on the instantiation of RA(HCPS) for this application presented in Section 2 in the companion paper Part II [3]. This case study highlights “small” failures (miscommunication, mishearing, violations of beliefs and expectations) in a fully functional situation, which can be problematic but need not necessarily be safety-critical. We illustrate established methods for system safety assessment, anchored as standard in the Aircraft Recommended Practices ARP 4574 [17] and called **preliminary hazard analysis** (**PHA**) [19] and **Hazard Operational Analysis** (**HAZOP**) [17]. We show that when these established methods are applied to instantiations of RA(HCPS), they can identify hazardous instances of wrong or inconsistent beliefs due to lack of communication within the cockpit system, which can lead to a failure and even a crash. This method is based purely on the structure of a system, such as given by the instantiation of the reference architecture to the cockpit interaction, and thus requires no deep semantical foundations. However, other methods are required in later stages of the safety and system development process which demand a semantic foundation, such as provided in the Section 3 of this article. This section thus serves as an “appetizer” fostering an intuitive understanding, in highlighting the additional hazards coming from misconceptions of beliefs among team-players controlling jointly safety critical processes, as made explicit in instantiating the reference architecture.

In developing a commercial aircraft, achieving high safety levels is of utmost importance. Aviation accidents can be traced to a variety of causes: pilot error, air traffic control error, design and manufacturer defects, maintenance failures, sabotage, inclement weather, among other tragedies [18]. Applying early safety analysis methods such as PHA [19] and HAZOP [12] as prescribed by

ARP 4574 [17] on instantiations of the reference architecture to model the cockpit assures that all hazards resulting from inconsistent beliefs and misconceptions are addressed, because these are first-class citizens of the reference architecture. In this section we illustrate this by first applying PHA and then HAZOP on the instantiation of the cockpit metamodel from the companion article Part II, thus analyzing the extent to which the miscommunications observed in the case study affect the health of the aircraft system.

2.1 Preliminary Hazard Analysis (PHA)

The PHA [19] serves as a first attempt at the safety evaluation process to categorize potential hazards by assessing the risk early in the development process. Distorted perception, limited direct observability, and distorted communication can cause imperfect beliefs; compromised information results in mistakes that jeopardize the integrity of the system.

An ego's incorrect perceptions can impact beliefs on all levels of the metamodel. Particularly when compounded, the actions that result from imperfect beliefs may endanger the integrity of the aircraft. Through system interaction between the pilots, the autopilot system, as well as all other communicating parties, there is a pressing risk of mode confusion. This sample PHA is not striving to be complete, but rather to exemplify how PHA can be applied to certain misunderstandings that are based on belief inconsistencies. Tables 1 and 2 recall well known classification schemes for establishing risk classes and the ALARP principle of required levels of risk reduction.

Risk Matrix

Requirement: Risk level must be **As Low As Reasonably Possible** (**ALARP** principle)

Table 1. Risk Matrix used for PHA

Frequency/ Consequence	1- Very Unlikely	2 - Remote	3 - Occasional	4 - Probable	5 - Frequent
Catastrophic	🟨	🟥	🟥	🟥	🟥
Critical	🟩	🟨	🟨	🟥	🟥
Major	🟩	🟩	🟨	🟨	🟥
Minor	🟩	🟩	🟩	🟨	🟨

Risk Levels and Actions

Table 2. Risk Levels and Actions used for the Classification of Hazards

Level	Name	Description
H (Color code red)	High	High risk, not acceptable. Further analysis should be performed to give a better estimate of the risk. If this analysis still shows unacceptable or medium risk, redesign or other changes should be introduced to reduce the criticality.
M (Color code yellow)	Medium	The risk may be acceptable, but redesign or other changes should be considered if reasonably practical. Further analysis should be performed to give a better estimate of the risk. When assessing the need for remedial actions, the number of events of this risk level should be taken into account.
L (Color code green)	Low	The risk is low and further risk reducing measures are not required.

An example of the application of the PHA to a subsystem is given by Table 3 using the indices and nomenclature of the aforementioned framework. The operating mode for the study is climbing, and the analysis below discusses potential hazards of the aircraft radio subsystem. Some of the evaluated hypothetical scenarios are actualized in the context of the sample study, and their severity levels are assessed here. It is important to recognize that an aggregation of multiple hazards, even those deemed minimal risk, can result in a more severe system failure. The Air France 447 crash is a real-world example of compounded miscommunications that resulted in catastrophe. When airborne in the midst of an unexpected weather event, the pitot tube froze due to inclement conditions which resulted in the disengagement of the autopilot system without explanation. The conditions necessitated manual operation; due to conflicting diagnoses of the malfunction, the pilots conducted incompatible emergency procedures. As a result of the inconsistent cognition, the aircraft entered a stall, then crashed into the sea at 200 km/h [7].

Consider the following excerpt of a corresponding PHA analysis. It shows multiple situations where wrong or inconsistent beliefs can lead to major or even critical hazards:

- Ref. 2.1. Failure of communication with other aircraft or tower will lead to inconsistencies of beliefs about cleared height levels, potentially causing inadvertent intrusion into the safety envelope of other aircraft.
- Ref. 2.4.1. When a crew is flying together for the first time, the likelihood that communication between pilots is misinterpreted is probable, e.g., due to cultural differences. Such a misinterpretation, e.g., the failure to point out a missed confirmation of the newly granted flight level, can lead to inconsistent beliefs between the tower and the captain.
- Ref. 2.5. The pilot misses a command or a portion of a directive, leading to a discrepancy of expectations of the correct operations of the aircraft. In practice, this mistake is often mitigated by redundant readbacks, but it can potentially compromise safety if repetition is omitted, misheard again, or otherwise wrongly interpreted due to negligence.
- Ref. 2.6. The pilot interprets misinformation via either an acoustic error (the pilot believes he hears a different command than what is correct) or a processing error (the pilot hears the correct information but associates with it a false meaning). The consequences of the latter may be more severe, as individual mental information processes lack the assurances that the cockpit's redundant communicated information does. Communication in this system is continually observed, assessed, and verified by the communicating agents and the observing participants. Such belief imbalances can manifest in radio communication as observed in the case example from partner paper Part II where the captain is guilty of all infractions listed as Ref. 2.6.1, Ref. 2.9, and Ref. 2.11.

System: Aircraft Radio
Operating Mode: Climbing

Table 3. A sample PHA Analysis based on the Instantiation of the Reference Architecture for Cockpit Interaction

Ref.	Hazard	Accidental Event (what, where, when)	Probable causes	Contingencies/ Preventative Actions	Prob.	Sev.	Risk Level
2.1	Communication with other aircraft or tower is impossible	Communication with other aircraft or tower is impossible in a trafficked airspace	Instrument malfunction (blown fuse, malicious attacks, extreme weather event)	Multiple radios available, emergency transponder number (squawk 7600) can be input to signal aircraft lost communications and be directed using aviation light signals	1	Critical	H
2.2	Communication with other aircraft or tower is limited or obstructed	Communication with other aircraft or tower is limited in a trafficked airspace	Interference (5G towers), break, weak point, loose connection in wiring between antenna and tuner	Multiple radios available, emergency transponder number (squawk 7600) can be input to signal aircraft lost communications and be directed using aviation light signals	2	Major	M

(Continued)

Table 3. Continued

Ref.	Hazard	Accidental Event (what, where, when)	Probable causes	Contingencies/ Preventative Actions	Prob.	Sev.	Risk Level
2.3	Communication between one or more members of the cockpit is impossible	Communication between one or more members of the cockpit is impossible during flight	Malfunction of aircraft radio system, subsystem, or accessory technologies such as squelch control (eliminates background noise)	Contact ATC regarding issue and seek instruction on how to best mitigate	1	Major	M
2.4	Communication between one or more members of cockpit is limited	Communication between one or more members of the cockpit is technically limited or obstructed during flight	Malfunction of aircraft radio system, subsystem, or accessory technologies); malfunction of pilot's headset; incorrect/inefficient squelch control	Backup headsets; emergency radio; emergency transponder number code (squawk 7600)	2	Major	M
2.4.1		Communication between pilots is misinterpreted	Cultural differences; limited perception capabilities of other pilots; intonations lost in radio communication	Ask for clarification; introducing pilots earlier (over time, they work together more succinctly)	4	Major	H
2.5	Pilot misses ATC's instruction	Pilot misses a change of heading instruction during operations	Lack of experience or background knowledge to interpret ATC's meaning or formulaic radio structure; pilot mental/physical fatigue; pilot focus elsewhere or otherwise preoccupied; ATC's lack of clarity or rushed speech; technical/radio obstruction; expecting standard, but instead granted special instruction	Multiple pilots with shared access to information; ask for a repeat or clarification if unsure; standard to conduct readbacks to catch mistakes	3	Major	M
2.5.1		Pilot misses an instruction to change altitudes	Lack of experience or background knowledge to interpret ATC's meaning or formulaic radio structure; pilot mental/physical fatigue; pilot focus elsewhere or otherwise preoccupied; ATC's lack of clarity or rushed speech; technical/radio obstruction; expecting standard, but instead granted special instruction	Multiple pilots with shared access to information; ask for a repeat or clarification if unsure; standard to conduct readbacks to catch mistakes	3	Major	M
2.5.2		Pilot misses a specialized command during operations	Lack of experience or background knowledge to interpret ATC's meaning or formulaic radio structure; pilot mental/physical fatigue; pilot focus elsewhere or otherwise preoccupied; ATC's lack of clarity or rushed speech; technical/radio obstruction; expecting standard, but instead granted special instruction	Multiple pilots with shared access to information; ask for a repeat or clarification if unsure; standard to conduct readbacks to catch mistakes	3	Major	M
2.6	Pilot mishears ATC's instruction	Pilot mis-hears/misinterprets ATC's instruction of a certain heading	Lack of experience or background knowledge to interpret ATC's meaning or formulaic radio structure; pilot mental/physical fatigue; ATC's lack of clarity; technical/radio obstruction	Multiple pilots with shared access to information; ask for a repeat or clarification if unsure; standard to conduct readbacks to catch mistakes	4	Major	M

(Continued)

Table 3. Continued

Ref.	Hazard	Accidental Event (what, where, when)	Probable causes	Contingencies/ Preventative Actions	Prob.	Sev.	Risk Level
2.6.1		Pilot mishears/misinterprets ATC's instruction of a certain frequency change	Lack of experience or background knowledge to interpret ATC's meaning or formulaic radio structure; pilot mental/physical fatigue; ATC's lack of clarity; technical/radio obstruction	Multiple pilots with shared access to information; ask for a repeat or clarification if unsure; standard to conduct readbacks to catch mistakes; frequencies given through detailed digital or analog regional maps, FAR/AIM,[1] Flight Management programming	4	Major	M
2.7	Communication between pilot and passengers obstructed	Radio system for general announcements to passengers malfunctions	Interference, radio, or speaker malfunction	Multiple instruments with same projection capabilities; pilot can go to cabin and make general announcement; pilot can alert flight crew and have them manually disseminate important information	1	Minor	L
2.8	Incorrect frequency input	Desired frequency is incorrectly input into the radio system	ATC misspoke; pilot misheard; pilot misread frequency from map; erroneous map	Multiple radio interfaces to store frequencies; multiple maps with frequencies outlined; automatic presets on software, i.e., FAR/AIM (updates yearly) and VFR (Visual Flight Rules) map/aeronautical chart (regional)	4	Major	M
2.9	Miscommunication/ misunderstanding of ATC's instruction	Pilot misunderstands ATC's instruction and operates under false pretenses	Lack of redundant features; pilot fatigue	Multiple pilots monitoring and communicating	3	Major	M
2.10	Erroneous readback of ATC's instruction	Pilot either misunderstands or misspeaks during readback	Pilot mishears instruction	ATC or copilots correct false reading	5	Minor	L
2.11	Pilot fails to conduct readback	Pilot neglects to readback information	Pilot mishears instruction	ATC or copilots correct false reading	5	Minor	L

Observing a system's distributed cognitive dissonance as a result of information transmission errors, as exemplified by this sample PHA, brings attention to new risk management strategies and highlights the key importance of design principles for Human-Machine Interaction outlined in the companion paper Part II. It is evident that the necessity of such risk management strategies can only be discovered and their sufficiency analyzed in a behavioral model permitting their description. This provides the rationale for including them and their causes in our semantic reference model.

2.2 Hazard and Operability Study (HAZOP)

The HAZOP methodology [12, 13] is a systematic team-based technique that can be used to effectively identify and analyze the risks of potentially hazardous process operations. This particular

[1] Federal Aviation Regulations/Aeronautical Information Manual, see https://www.faraim.org/

Table 4. Guide Words and Definitions for the HAZOP Analysis of the Cockpit Systems

Guide-word	**Meaning**
No (not, none)	None of the design intent is achieved
More (more of, higher)	Quantitative increase in a parameter
Less (less of, lower)	Quantitative decrease in a parameter
As well as (more than)	An additional activity occurs
Part of	Only some of the design intention is achieved
Reverse	Logical opposite of the design intention occurs
Early/late	The timing is different from the intention
Before/after	The step (or part of it) is affected out of sequence
Faster/slower	The step is done/not done with the right timing

hazard assessment only serves to illustrate the application of this hazard analysis method to instantiations of the reference architecture and is not in any way attempting to be complete. Whereas the PHA approaches analysis by system decomposition, HAZOP instead explicates what can go wrong in the communication between systems. A HAZOP focuses more particularly on system information flow which directly pertains to the function of this article.

The mere mental processing of commands can be considered as a series of subprocesses in which we perceive and generate mental representations which then influence beliefs. Communication and cognition are information pathways that can be distorted, falling prey to misinterpretation. As such, included in this HAZOP is an analysis of the information processing of a human entity; a person generates mental representations from words that are perceived and understood. These mental representations are highly dependent on the individual knowing certain ontologies with which to classify an understood element of information. Therefore, we determine the internal structure of a human to be comprised of (at least) two parts − one which is responsible for word recognition, and one which is responsible for mapping words into ontology. As such, an entire classification of failures can be caused by issues with either system process. A person can either not possess a knowledge of a certain word, or connect a word to an erroneous assumed meaning. The accident in these cases would be not in the acoustic perception of the correct recognition of a word, but a flawed semantic interpretation. This approach is exemplified in the following sample HAZOP.

Using the guidewords detailed in Table 4, an example of the application of the HAZOP analysis to a subsystem is given by the following table focusing on aircraft communications. The operational mode for the entire analysis is considered to be climbing, and the design intent is to transmit information timely, effectively, and efficiently with minimal technical or human error. The HAZOP allows for a more comprehensive analysis of component shortcomings within a system. For the purpose of the case study, the radio and adjacent communications were deemed the most relevant, and a series of hypothetical and real events are included in the HAZOP. Deviations, possible causes, and consequences are included as most likely to occur under the given circumstances, however, limitations of unexpected circumstances should be acknowledged.

Consider now the following excerpt of a HAZOP analysis:

- No. 1. Only part of the ATC-Aircraft communication was conveyed successfully. Failure of information distribution with other aircraft or tower can lead to inconsistencies of beliefs on the aircraft height, radio frequency, or instruction. The consequences of these misunderstandings range from marginal to severe depending on the infraction.

- No. 3. Before receiving instruction, a seasoned pilot may instinctively maneuver the aircraft due to a high confidence of expectation. In the event that the assumption is wrong, the expectations will be violated and could compromise the safety of the system.
- No. 4. The pilot does not conduct a complete readback. ATC assumes that the pilot either understands but is purposefully incomplete, or that the pilot missed information. In the event of an incorrect belief, consequences could be severe.
- No. 6. Pilots use abbreviations, colloquial language, or social references to communicate. Meaning is derived from the receiver of the message, a process which is highly dependent on ethnography, context, and social cues, rendering it highly vulnerable to misinterpretation.
- No. 7. The pilot inputs the incorrect frequency because of mishearing, misunderstanding, or perhaps a lapse in dexterity. This error results in an entirely incorrect new mode of communication.

System: Aircraft Communications

Design Intent: To transmit information timely, effectively, and efficiently with minimal technical or human error.

No.	Guide Word	Element	Deviation	Possible Causes	Consequences	Safeguards	Actions required	Action allocated to
1	PART OF	Radio	ATC operator misinterprets aircraft's intentions	ATC misspeaks; pilot misspeaks; ATC misunderstands; pilot misunderstands	Misunderstanding and discrepancy between beliefs (interpreted action of aircraft) and perceptions (data in front of them showing otherwise)	Multiple pilots listening; multiple ATC operators listening; advanced instrumentation detailing position, speed, and elevation within airspace	Explicit communication with sufficient readbacks between ATC and pilots	Pilot, ATC, ATC Instrumentation
2	TOO LATE	Radio	Late transmission	Aircraft out of range; ATC must repeat due to misunderstanding; interference; break; weak point; loose connection in wiring between antenna and tuner	Misunderstanding; collision in extreme scenario	Multiple frequencies	Pilot must switch frequencies depending on the location of the aircraft; ATC instructs to change frequencies	Pilot, ATC
3	BEFORE	Radio	Premature pilot maneuvers before authoriza-tion	Reflex; relying too heavily on expectations	Inadvertent violations of corporate or federal regulations	Two pilots in command, one supervising pilot not flying; corporate and federal regulations; mutually understood courtesy	Operate according to beliefs without relying too heavily on expectations that create bias	First Officer
4	PART OF	Radio	Pilot reads back only part of command	Laziness; efficiency; did not hear command	Pilot may miss critical instruction or have misunderstanding due to erroneous perception	Legal and company policy required readback of runways and other safety critical instructions; other pilots act as safeguards	Thorough and complete readbacks of ATC commands necessary for clear communication	Captain, ATC
5	NO	Radio	Pilot does not read back a command	Laziness; efficiency; did not hear command	Pilot may miss critical instruction or have misunderstanding due to erroneous perception	Legal and company policy required readback of runways and other safety critical instructions	Thorough and complete readbacks of ATC commands necessary for clear communication; ask copilots for clarification	Captain, ATC
6	LESS	Radio	Pilots do not explicitly verbally communicate with each other	Efficiency; social understanding; ethnographic grounding	Pilots may misunderstand each other because they are not explicit; intuition differs culturally and based on an individual's experience	Pilots should be concise but clear; role specific procedures to enable focus and attempt to not overwhelm	Pilot introduction, conversation, and practice, so that they can get accustomed to how they operate	Pilots

No.	Guide Word	Element	Deviation	Possible Causes	Consequences	Safeguards	Actions required	Action allocated to
7	NO	Radio	Incorrect frequency input	ATC misspeaks; pilot misheard; pilot misread frequency from map; erroneous map; pilot presses wrong number inadvertently	Lost communication with desired ATC; out of range of inputted frequency (hear nothing)	Radio calls begin with who aircraft is addressing (ex: Oakland Center) which can signal an erroneous frequency; multiple radios; read backs; FAR/AIM/Region-al maps/Flight Management System (FMS); company procedure	Entering tentative radio frequency for switch-over; read back	Captain

This analysis can be extended to any partaking system within the cockpit, most notably the advanced autonomous technical systems and the human participants. A HAZOP can identify errors in the transmission of information between entities. This proves to be a valuable tool in developing safety-critical human cyber-physical systems, as probabilities associated with system errors can be assigned to hypothesized errors, resulting in proper risk mitigation in the preliminary design process. A hazard analysis approach adapted to the reference architecture cockpit sample case provides a unique focus on information comprehension, distribution, and cognition – all of which have been historically responsible for a large portion of system failures. This example renders evident the necessity of probabilistic models of belief not only about ground-truth state (which may be uncertain due to inexact measurements and partial observability as well as uncertain dynamics), but also concerning human state and comprehension. Again, this provokes the quest for a behavioral model permitting such descriptions, providing the rationale for including them into our semantic reference model.

3 COMPOSITION AND SEMANTIC FOUNDATION: INITIAL THOUGHTS

The underlying semantic model of a system will be built on dense time probabilistic branching structures and games on these with imperfect information on a state space which is reflecting states and beliefs of all systems. Branching situations reflect decision points, where the ego system´s reaction is chosen following its strategy to react to new beliefs about its environment, or internal non-determinism, or probabilities coming from the use of probabilistic models to predict environment dynamics.

This section first elaborates a trace-based semantics for single systems, then introduces the novel classes of games required to naturally model the targeted application classes, and then defines the semantics of interactions between such games.

3.1 The Semantic Baseline

We assume that all systems are instances of a (possibly countable) set of classes C. In any practical application, a taxonomy will define what types of systems are relevant and define their relevant attributes. Typically, this taxonomy is equipped with a partial order. If $C \leq C'$, then C inherits from C' all its attributes and either specializes these or extends these, and the dynamic capabilities of C are contained in those of C'. E.g., C'=*vehicle* would specialize both to C=*truck* and C=*car*, and the superclass *vehicle* would allow for any behavior of its possible instantiations.

All instances of the same class have the same set of states, set of roles, set of observables, the same perception capabilities, the same dynamic capabilities, and the same set of actions.

Actions range from those controlling the dynamics of the system, to forming and leaving coalitions with other systems, to incorporating other systems.

Systems can be stand alone or part of an aggregate system. Systems can be created and destroyed.

All instances of systems have a unique identity i drawn from a countable set of identifiers I. We denote the class of a system with identity i by $i.class \in C$. We write $i \in i'$ to denote that system i is a component of an aggregated system i' *and* denote the set of identities of all subsystems of an aggregated system i' by $i'.ss$.

A trace describes for every instance i of a system at any point in time the current valuation of

(1) its *state i.s*, including as special states *unborn, alive, dead*
(2) its health state i.health
(3) its current *role i.role*
(4) its current set of *goals i.goals*
(5) its current partial order i.≤ of its current set of goals i.goals
(6) its current dynamic capabilities i.dyn
(7) its current set of possible *actions i.act*
(8) the current set of *systems perceived in its environment i.env*, identified by a set of names picked from a countable set *i.envI*
(9) its current set of *beliefs* about currently perceived systems in its environment *i.beliefs*
(10) its current set of coalition partners i.partners, a subset of i.env

In differentiating *roles* from *states*, as proposed in [6], we intend to model the differences between continuously evolving aspects of the system on one side, and discrete changes from one role to another role, which typically remain stable for significant time intervals. Role changes can be both induced from changed beliefs (e.g., about the health state of the patient, switching from performing surgery to emergency measures to stabilize the health state of the patient) or deliberate acts of the ego system. Role changes typically come with changes of goals, or changes of the importance of goals as reflected in their partial order. Role changes typically also involve a change of dynamic capabilities and available actions. They may also influence the set of current beliefs, because the attention is now addressed to perform possibly completely different tasks.

Let us elaborate the pragmatics behind these definitions using examples.

Example 1 (Autonomous Driving).

> Car *bad* of class *ROBOTCAR* is driving on a multi-lane street at night. All its components, including a stereo-video camera and the perception subsystem, as well as all subsystems controlling the car dynamics are in good health state, hence *bad.health = excellent*. Car *bad* perceives at time t_0 some object ahead at an approximate distance of 120 meters. Hence it creates a new identifier *something* from its set of environment identifiers *bad.envI*. It initializes beliefs about something, in the form of distributions over the identifier properties. For example, characterizing the belief over distance between *bad* and *something* as a Gaussian distribution, $d(bad, something)\ N(\mu = 115, \sigma = 5)$, and the belief over the object class as a categorical distribution, $\mathrm{P}(something.class = bicycle = 0.8, P(something.class = pedestrian) = 0.2$. At time t_1 the perception system measures *something* to be of class *pedestrian*, and the distance to *something* to be 75 meters. The beliefs on these quantities are updated by computing the appropriate posterior distributions, resulting in higher certainty beliefs about the state of *something*. Hence the control of car *bad* decelerates the vehicle since it contains the goal to avoid accidents with pedestrians. However, at time t_2 the perception system updates again the belief about something, now considering this to be, with high probability, a mere shopping bag. Even though the belief about the distance is now updated to be very small with high probability, the car resumes full speed.

The corresponding *ground truth observations* of *something* at times t_0, t_1, t_2 are

(1) something.id= Jane Wilhelmson, Jane Wilhelmson.class =pedestrian, Jane Wilhelmson.state = alive
(2) something.id= Jane Wilhelmson, Jane Wilhelmson.class =pedestrian, Jane Wilhelmson.state = alive
(3) something.id= Jane Wilhelmson, Jane Wilhelmson.class =pedestrian, Jane Wilhelmson.state = dead

The overarching quality criteria for the sensor and perception system demands at all times **that the distance between beliefs about systems in the environment of the ego-system and the corresponding ground-truth is to bound the criticality of misperceptions**. Although knowing the identity of the sensed something is irrelevant for determining the trajectory of *bad*, the incorrect classification is a major violation of the demanded quality of environment perception, resulting in the death of the person crossing the street.

Example 2 (Cooperative Driving).

> Car *clever* of class *CoopCAR* drives on a country road. When just having passed a curve, it perceives an object in its lane about 40 meters ahead, so close that even while emergency breaking an impact cannot be avoided. Car *clever* perceives a vehicle *somevehicle* approaching at the opposite lane, with a current distance believed to be about 250 m and unknown speed. Car *clever* decides to try to form a coalition with *somevehicle*, while at the same time initiating an emergency braking maneuver. Luckily, *somevehicle* is also equipped with car-to-car communication and responds to the distress message of car *clever* by acknowledging to accept as its top priority goal a drastic reduction of its own speed. At the same time, *clever* and *somevehicle* exchange their beliefs about *something*, leading to belief revision with a high value of *P(something.class=human).* The message is received in time by car *clever* to initiate a lane change before impacting the human body. After safely passing the human body, *clever* initiates an emergency rescue call informing about the exact position of the human body. The call is picked up by all cars in the vicinity as well as by the nearest rescue center, which dispatches a rescue helicopter.

Through forming a coalition, *clever* thus achieved sufficiently precise information about the speed of *somevehicle* and the class of *something* to infer that avoiding an impact has top priority and that this can be achieved by a lane change.

We now turn toward the formal definition of traces. Each trace subsumes the ground-truth observations of all alive systems. Each trace extends these by the beliefs of all alive systems.

$$\text{tr}: \mathfrak{R} \rightarrow \{ < \text{i, i. configuration} > \mid \text{i} \in \mathbf{I} \wedge \text{i. s} \mid = \text{alive}\}$$

where
i.configuration ∈
[[i.C.states]] — its current state
x *[[i.C.health]]* — its current health state
x *[[i.C.roles]]* — its current role
x *[[i.C.goals]] x ([[i.C.goals]] x[[i.C.goals]] →{0,1})* — its current partially ordered set of goals
x *[[i.C.dyn]]* — its current dynamic capabilities
x *i.envI* — its current set of names for currently perceived systems
x *{<i′, i′.configuration >| i′∈ i.envI}* — its current beliefs about such systems
x *i.envI* — its current set of coalition partners
x *℘(i.envI)* — its current set of subsystems
x *℘(i.envI)* — its current set of supersystems

Here we use *[[S]]* to denote the semantic domain associated with syntactic category *S*. Since these are aggregation level dependent, the concrete definition of these is not further elaborated in this article.

Note that the above definition is necessarily recursive, because of the need to model beliefs of beliefs of .... of beliefs.

3.2 A Game-Theoretic Semantics of Ego Systems

To capture the interaction between an ego system and its environment (including other system instances, whether they are human, environmental, or otherwise), we leverage a game-theoretic interpretation. We assume that ego systems exist in the context of a *game instance G*. Specifically, a game G is played among a *set of players* $\boldsymbol{P}$, where each player $P \in \boldsymbol{P}$ is a player instance. Player instances are simply instances of systems as described in the previous section, with a few additional properties relating to the game they are a part of. In what follows, we formally define a game, a player, and related concepts.

To represent interaction among players, a game instance G is associated with the following properties:

(a) Its *set of players,* $G.\boldsymbol{P}$, where the notion of a *player* is defined below,
(b) Its *duration,* $G.d$, which can be finite or infinite,
(c) Its *playable domain,* $G.D$.

Each player P within the set of players $G.\boldsymbol{P}$ is defined by the attributes

(1) Its *system instance P.sys*, as defined in previous section
(2) Its *decision process P.dec*, which is its method for determining actions.

It is assumed that all games are played in continuous time, for the duration $G.d$. Throughout the game duration, the players' decision processes generate actions to play and times to play them. The decision process for each player can be any process that maps from game traces to actions. For example, a decision process could be a pre-specified and fixed sequence of actions, a generator of random actions according to a specified distribution, a human decision process, or a process to rationally generate actions according to some decision model. For some players, the decision model itself might be represented by an internal model of the game, specified as a separate instance G'. This game instance may change to match the current beliefs of the owning player, about the instances of all other players in the game, and may use a simplified duration or domain compared to the actual game instance G.

To concretize the above definitions, consider the following example.

Example 3 (Unmanned Aerial Vehicle).

> An **Unmanned Aerial Vehicle** (**UAV**) is considered the ego system of interest. The UAV is assumed to operate in a shared airspace with another human-piloted aircraft. There is considerable strong wind acting in the airspace. To model the interaction between the two aircraft as well as the environmental factors (wind), a game instance G is created. The playable game domain is the airspace of consideration, which is specified as a geofenced volume. The game duration is set as the longer value of the maximum possible flight times of the two aircraft. The set of players $G.\boldsymbol{P}$ is the set consisting of the UAV, the human-piloted aircraft, and a third player representing the environmental wind factors.
> The environmental player is represented by a system instance of class *wind*. The instance properties for the *wind* class can still be specified in the semantics outlined in the previous section, although many of the properties may have trivial or null values. For example, the

> *wind* instance *role* is *disturbance*, the dynamic capabilities are to apply forces upon the airspace and aircraft within, and the set of possible actions are the set of all possible vector fields representing the airflow in the airspace. The *decision process* for this player is specified by a random generator according to an appropriate probabilistic weather model.

Both aircraft players are systems represented by instances of class *aircraft*, with associated *state, health, role, goals*, and so on. as outlined in the previous section. The *decision process* for the human-piloted player is a *human decision process*, whereas the UAV player's decision process is represented by a *game-theoretic decision process*. Specifically, the UAV uses its current beliefs at any time *t* to formulate a simplified game aimed at modeling the interaction among the UAV, the human-piloted aircraft, and the environment. The simplified game *G'* also contains a set of three players, representing the three players in the actual game *G*, but with estimated system instances and decision processes. For example, in the game instance *G'*, the environmental *wind* player as well as the human-piloted player may be assumed to have an *adversarial decision process*, which generates actions that work counter to the UAV's goals. Similarly, the game duration for *G'* may be of a shorter time horizon to simplify the considered interaction. The UAV's decisions at time *t* are generated by computing equilibrium strategies for the instance *G'*.

In the UAV example, the ego system's decision process was represented as a procedure that computes *equilibrium strategies* of a privately held game instance, which serves as a model of the true game being played. Although there are multiple possible definitions of equilibrium in this context, we broadly define an equilibrium to be a set of actions for all players, spanning the duration of the game, such that no player can improve upon their objective by changing their actions, given the restrictions on their set of possible actions, and an assumed *information structure* of the game.

The notion of an *information structure* is central to the definition of an equilibrium and is precisely what distinguishes the classic concepts of equilibrium from one another (Nash, Stackelberg, and so on.; see references below). An *information structure* defines precisely how much information is available to players when defining what their rational or optimal decisions are. To illustrate this concept, consider the following example.

Example 4 (Information Structures – Autonomous Shepherd).

> Consider a game *G* played between two systems: a *shepherd* and a *flock of sheep*. In this example, the *shepherd* aims to herd the flock into a designated corral, with the entire field being the domain of the game. We will ignore any environmental factors in the game, which could be represented as additional players in the game. The game played between the shepherd and flock has a finite duration, lasting from the start of the game until sundown or until the flock has been successfully corralled, whichever occurs first. Similar to the previous example, the shepherd may generate a representative game instance *G'* as part of its decision process. In this example *G'*, the flock is modeled as a system which contains a set of attributes as defined in the previous section, including a simple objective which aims at staying close to the shepherd and otherwise exert minimal energy. The shepherd's modeled objective in *G'* is, as in *G*, to corral the flock from its initial configuration into the designated location, and otherwise use minimal energy. A shortened or simplified duration of the game may be considered in *G'*.

When determining an equilibrium strategy for the game *G'*, such as to be used by the *shepherd* in *G* to make decisions, an *information structure* must be assumed. One such possible structure is a *flat* structure, which corresponds to a *Nash equilibrium*. In this flat structure, an equilibrium of *G'* is one in which actions for both the *shepherd* and the *flock* are optimal *without considering any reaction* *of each other to one's change in actions*. In particular, the resulting equilibrium strategy is for the shepherd to not move at all and for the flock to gravitate toward the shepherd's current location. To see why this is, consider the action of the shepherd. The equilibrium strategy is to not move at all, which is optimal with respect to the objective of using minimal energy but, unless the shepherd is already in the corral, suboptimal with respect to the objective of herding the flock. However, since for the given information structure the shepherd cannot reason that if he moves, the flock will follow, any other action will simply increase the cost with respect to the minimal energy goals and will have no effect on the herding goal. Therefore, the strategy to stay put is optimal.

On the contrary, using a more sophisticated information structure, the equilibrium strategy for the shepherd will achieve the expected behavior of moving to the corral, so that the flock of sheep follows. One example structure which achieves this result is the *leader-follower structure*, which will result in the classical *Stackelberg equilibrium*. In this structure, one player takes the *role* of the *leader*, and the other the *follower*. The *leader* may anticipate the reaction of the *follower* when considering what constitutes an optimal strategy, whereas the *follower* acts as in the flat information structure. Assuming the *shepherd* is the *leader*, it can reason that moving from its initial configuration will induce the flock to follow, and the optimality of moving to the corral is established.

In the given examples so far, the privately held game instances *G'* which model the true game *G*, and which are used to compute decisions, technically also assume a decision process for each of the players. In the context of using these game instances to compute equilibrium strategies, the resulting decision processes are assumed to simply be optimal under the given information pattern, meaning they achieve the corresponding equilibrium.

We can use the semantics of games and information structures to model collaboration among systems as well.

Example 5 (Collaborative Driving).

> Consider an autonomous vehicle (*ego* system) which is driving on a two-lane road at high speeds. Upon rounding a corner, the ego system detects an obstacle obstructing the lane in which it is driving, as well as another vehicle (*other* system) driving in the oncoming lane. There is not enough time to stop to avoid hitting the obstacle obstructing the lane, and furthermore, the shoulder of the road on the ego-lane side is nonexistent. However, there is room for the oncoming vehicle to pull off onto the shoulder on its side of the road, allowing for the ego system to temporarily swerve into the oncoming lane and avoid the obstacle, and both vehicles are equipped with vehicle-to-vehicle communication capabilities. The formation of the coalition necessary to avoid collision is naturally modeled in the game theoretic semantics defined in this section.
>
> This situation is captured as a game played between the two vehicles (*ego* and *other*), in a playable domain which is the road segment they share, and immediate surrounding area (road shoulders, etc.). The game duration is the time from which both vehicles have entered the road segment vicinity, to the time when either agent leaves the domain. Both player instances have objectives of maintaining safety, the rules of the road, and a comfortable driving profile. The decision process for the ego system is, as in previous examples, defined in terms of a privately held game instance *G'* and a corresponding information structure defining equilibrium strategies for the agents in *G'*. The game instance *G'* (which may be constantly updated as the ego system updates its beliefs about the true game *G*) includes an accurate model of both system's capabilities and objectives, as is reasonable in this situation. At the onset of game *G*, the ego system has just detected the obstacle and oncoming vehicle but has not established a coalition with the oncoming vehicle to ensure cooperation of the safety maneuver. This coalition emerges as rational play of the game under

an appropriate information structure, which in this case is a *feedback structure* – meaning rational decisions reason about actions in the present will influence both agents' decisions in the future.

At the onset of the game, the ego agent may initiate a request to the oncoming vehicle to engage in a collaborative maneuver but, until the request is approved, cannot assume with full confidence that the other vehicle will act cooperatively. A probability is assigned to the likelihood that the other agent will respond affirmatively to the request for cooperation. Because a request for cooperation is effectively a zero-cost action, it emerges as a rational strategy to immediately employ. The rational *physical strategy* of the ego vehicle is to stay in its lane on a collision course with the obstacle. Although this is an unsafe action, it is preferable to swerving into the oncoming vehicle, which is both unsafe and violates the rules of the road.

While the ego system is executing this strategy, the other system approves the request for cooperation. This leads the ego system to update its belief about the preferences of the other system to reflect that it now prioritizes the safety of the ego system over its own satisfaction of the rules of the road. The game instance *G'* now admits strategies for both agents in which the cooperative evasion maneuver is rational for both agents. Note that a feedback information structure is essential for this behavior to emerge. Because the other agent must be able to reason that by driving on the shoulder of its own lane, it will create space for the ego system to maneuver into its lane at future stages, and it will lead to the desirable safety of the ego system.

The concept of privately held game instances is closely related to that of *incomplete information* in extensive form games, as discussed by [11]. Games played with incomplete information are used to specify interactions in which players are not certain of the underlying capabilities or intentions of other players in a game. These games can be transformed into games of imperfect information [16], in which the incomplete knowledge of the game being played can be expressed as unobserved actions chosen by an auxiliary "nature" player, responsible for choosing which possible reality is true.

The decision processes associated with each player (as described in this section) may be represented by privately held game instances, and these game instances may themselves have incomplete or imperfect information. With an associated information structure (alternatively, *information pattern* [2]) a notion of rationality is implied, thereby defining a notion of equilibrium and ultimately the decision process for that player. To the best of our knowledge, no existing work considers the interaction of players who play rationally according to their own privately held, independent game instances. A formal semantics for modeling the structure of this interaction is provided in the section to follow.

3.3 Interaction Semantics

The previous two sections have introduced a generic model of agent state including (fallible) beliefs about other agents and the environment as well as information frames permitting an agent to locally construct rational strategies based on local beliefs about the structure of the game arena and the states and strategies of other parties. Based on these, we are now progressing to exposing the interaction semantics describing the *emergent joint behavior of interacting agents.*

A fundamental consideration underlying the proposed semantics is that belief revision and local rational strategy construction may interact non-trivially, given that the —believed to be true— local copy of the game arena may become subject to belief revision based on observations of other agents and their and the environment's actual dynamics (i.e., based on observation of parts of the ground-truth arena). Such a dynamic belief revision concerning the arena underlying the local strategy selection may obviously induce a dynamic local strategy revision, such that we must render game arenas as well as local strategies first-class citizens of agent state, enabling the encoding of their mutual revisions as transitions.

This implies that the state of an ensemble of interacting agents is described as a set $\{(x_1, s_1), \ldots, (x_n, s_n)\}$ of pairs of agent states x_i and current agent strategies s_i, with the latter resolving any choices in the agent's local transition system. The individual agent states x_i are type-consistent mappings $x_i : V_i \rightarrow v \in V_i D_v$, where D_v is the domain of variable v. We assume in the sequel that the variable name space of individual agents is disjoint, i.e., $\forall i \neq j : V_i \cap V_j = \emptyset$.

The local transition system of an agent i thereby features four kinds of transitions:

(1) *Time passage transitions* governed by uninterrupted differential dynamics: for any $t \in R_0$, the notation

$$x_i \rightarrow_i^{(t,u)} X'_i,$$

denotes that state x_i evolves to state distribution X'_i by an uninterrupted continuous evolution of duration t following the continuous dynamics associated to agent i (which in general is defined by some form of differential equation) under type-consistent random input

$$u : [0, t] \rightarrow \prod_{j \neq i} D_j,$$

where D_j denotes a distribution over type-consistent state valuations $V_j \rightarrow v \in V_j D_v$.

Note that such continuous behavior will in general also include dynamic updates of beliefs according to the known or believed dynamics of other agents.

Example 6

As an example, consider an agent believing car *red* to be at position $(x, y) = (10.7\text{m}, 20.2\text{m})$ and to move with constant speed

$$\left(\frac{dx}{dt}, \frac{dy}{dt}\right) \approx \left(U_{[10, 10.8]}\text{m/s}, 0\right),$$

where $U_{[l,u]}$ denotes the uniform distribution over $[l, u]$. After a duration of t seconds, the agent will believe that car *red* is positioned at

$$(x, y) = \left(U_{[10.7+10t, 10.7+10.8t]}\text{m}, 20.2\text{m}\right),$$

leading to a continuous belief update.

Note that we are not imposing any specific assumptions on the type of continuous dynamics, which consequently may be given explicitly as a function of time passage, as an ordinary or algebraic differential equation, or may include retarded dynamics covered by, e.g., delay differential equations.

(2) *Instantaneous state updates* subject to random distributions: the notation

$$x_i \rightarrow_i^{\text{jump}} X'_i,$$

denotes that a discrete jump is enabled in state x_i and that x_i evolves to state distribution X'_i when the jump is taken.

Example 7

As an example, consider a noisy measurement process where a measurement is taken whenever the time variable reports time to be a multiple of the sampling interval and where the instantaneous state update noisily copies the value of a physical state variable x into a measurement m_x according to $m'_x := N(x, 0.1)$, providing for measurements normally distributed around the true value.

State updates can cover physical state (e.g., the state of the register holding the measurement m_x) as well as beliefs. The latter may e.g., happen when a coalition of agents is formed, and beliefs and knowledge are exchanged by messages within the coalition. In such cases, both the formation of the coalition as well as the knowledge exchange require a sequence of messages, which can individually be modeled as updates of the recipients' states.

(3) *Discrete decisions* governed by the currently active strategy: the notation $(x_i, s_i) \rightarrow_i^{\text{decide}} X'_i$ denotes that a decision is enabled in state x_i, which then is resolved according to the currently active strategy s_i of agent i such that $X'_i = s_i(x_i)$. Note that the latter allows for adopting randomized as well as pure (by means of Dirac distributions) strategies.

Example 8

As an example, consider an autonomous car leaving an intersection and facing the choice between two parallel lanes. This choice will be resolved by the current strategy, which may dictate using the right lane if it is empty (no matter what the state of the left lane is), but the left lane should the right lane be clogged.

(4) *Strategy updates*: the notation

$$(x_i, s_i) \rightarrow_i^{\text{SUpd}} s'_i,$$

denotes that a strategy update is found to be due in the states-strategy combination (x_i, s_i) and that such update leads to the new local strategy s'_i.

Example 9

As an example, consider that the ego car has observed other cars avoiding a crowded intersection by regularly taking a shortcut across private property. The observation that this seems to be accepted by the property owner induces a sequence of increasingly certain belief revisions concerning the route options being available to the ego car, which eventually prompts a strategy revision due to the consolidated difference between the actual set of options and those having been available when computing the rational strategy s_i.

Note that it might make sense to assume a minimal temporal distance Δt between any two strategy revisions or between the reason for strategy revision and the strategy revision actually being computed and implemented, which, however, can easily be encoded into the transition system without need for extra mechanisms.

Note that strategies have thus become first-class members of the state space, as have their updates based on transitions.

We now progress to defining the semantics of actual interaction between multiple agents. We herein consider discrete actions as being urgent,[2] arriving at the following transition kernel of the interaction semantics:

$$\{(x_1, s_1), \ldots, (x_n, s_n)\} \rightarrow \{(X'_1, s'_1), \ldots, (X'_n, s'_n)\},$$

iff one of the following conditions holds: *Time passage:*

$$\{(x_1, s_1), \ldots, (x_n, s_n)\} \rightarrow \{(X'_1, s'_1), \ldots, (X'_n, s'_n)\},$$

[2] This is a matter of mere convenience, as it saves us from introducing numerous extra concepts like invariants, which, however, can be added by standard semantic means should urgency be considered inadequate.

iff

- —for all agents i, the local dynamics is respected when considering the other agents' dynamic output as input, i.e., there is a type-consistent joint trajectory $traj : [0, t] \to jV_j \to jv \in V_jD_v$ that respects all agent-local dynamics due to
- - $traj(0) = \text{Dirac}(x_1 \oplus x_2 \oplus \ldots \oplus x_n)$, where $\text{Dirac}(x)$ denotes the Dirac distribution at x,
- - $x_i \xrightarrow[i]{(t, traj|_{j \neq i V_j})} X'_i$ for all i and each $t' \in [0, t]$,
- —no instantaneous actions in the form of a jump, a discrete decision, or a strategy update become enabled along the trajectory, i.e., for no agent $j \neq i$ there are $t' \in$ and intermediate states $x'_i \in \text{carrier}(traj(t'))$ and follow-up state distributions X''_i or follow-up strategies s''_i with $x'_i \xrightarrow[i]{\text{jump}} X''_i$ or $x'_i \xrightarrow[i]{\text{decide}} X''_i$ or $(x'_i, s_i) \xrightarrow[i]{\text{SUpd}} s''_i$, respectively,
- —all strategies remain unchanged, i.e., $s'_i = s_i$ for all agents i;

Decision taken:

$$\{(x_1, s_1), \ldots, (x_n, s_n)\} \to \{(X'_1, s'_1), \ldots, (X'_n, s'_n)\},$$

iff

- —no strategy update is due,[3] i.e., for no agent $j \neq i$ there is a follow-up strategy s''_i with $(x_i, s_i) \xrightarrow[i]{\text{SUpd}} s''_i$,
- —a decision is enabled in some agent i, i.e., $x_i \xrightarrow[i]{\text{decide}} X'_i$ holds,
- —all other agents hold their state, i.e., $X'_j = \text{Dirac}(x_j)$ for all $j \neq i$,
- —all strategies remain unchanged, i.e., $s'_i = s_i$ for all agents i;

Strategy update:

$$\{(x_1, s_1), \ldots, (x_n, s_n)\} \to \{(X'_1, s'_1), \ldots, (X'_n, s'_n)\},$$

iff

- —a strategy update is enabled in some agent i, i.e., $(x_i, s_i) \xrightarrow[i]{\text{SUpd}} s'_i$ holds,
- —all other agents hold their strategy, i.e., $s'_j = s_j$ for all $j \neq i$;
- —all agents hold their state, i.e., $X'_i = \text{Dirac}(x_i)$ for all i.

This transition kernel immediately induces the full interaction semantics by the usual Markov chain rule for finite-trace probabilities being induced by transition probabilities. This extends to (countably) infinite traces by the standard cylinder or cone construction.

Note that the above semantics admits Zeno behavior, only local existence of solutions, and similar anomalies, as usual in the domain of hybrid dynamics. Due to race conditions between simultaneously enabled strategy updates, it furthermore admits natural non-determinism in its interleaving semantics, which however remains irrelevant when immediate strategy updates based on a single observation are excluded and strategy updates instead follow durational phases of successive belief update.

When composing systems, say S_1 and S_2, the part of the ground truth view of S_1 is now seen from the ego-view perspective, and similarly for S_2.

Example 10

As an example, let S_1 be a highly automated vehicle V, and S_2 be the driver D. We can take different perspectives depending on the type of analysis we are interested in:

[3] Imminent strategy updates take priority over strategy-led decisions.

- If we want to understand how to perceive the ground truth about the driver and how to adapt strategies of the autonomous vehicle so that the driving style induces trust and acceptance, then we consider the ego-view of *V* and view *D* as part of the environment of the system. Then the challenge is to identify those states of the driver which are involved in generating trust and acceptance, and find sensors allowing us to generate beliefs about such states of *D*. The ego view of *V* will contain at all levels of the hierarchy control components which incorporate driver models to anticipate the effects of strategy generation, maneuver selection, and low- level control on trust and acceptance, and thus, e.g., communicate to the driver that it reduces its speed because it has achieved information about icy road conditions ahead. Such actions of *V* will influence the ground truth of *D* and –depending on the quality of the internalized driver model – either actually contribute to strengthening trust and acceptance, or potentially have negative repercussions. In experiments with test persons (allowing to provide ground truth) we tune the perception system, driving strategies, and driver model until we have reached high confidence in the success of the interaction between *V* and *D* to create trust and acceptance.
- At a later stage in design, we want to focus on the interaction of *V* with other participants in a given traffic scenario. We then form *ego(V||D)*, that is consider the vehicle with its driver as one entity, and are focusing on the correct situational awareness and control in the environment *ENV(V||D)*, where thus *V* and *D* join perception, beliefs, and control to determine the actions of *V||D*. At this point, it is key that previous design steps have achieved shared situational awareness of *V* and *D*: if beliefs of *V* and *D* about the ground truth of *ENV(V||D)* disagree, say the driver misses a car approaching on the left lane on the highway and *V* failed to raise the awareness of *D* to this critical situation, then this results into a crash. Similarly, if *V* fails to recognize the woman ahead due to bad light conditions and heavy rain, but *D* sees this woman, a crash can only be avoided if *D* can take over control. Only after experimental validation ensures that *V* and *D* have consistent beliefs, plans, and actions can we hide internal interfaces in *ego(V||D)* and view this as one system.

From this example, we can derive the following key observations for the parallel composition of two systems:

- Such a composition is only valid if *ego(V)* and *ego(D)* are *consistent*, i.e., their beliefs differ only in irrelevant aspects, and their selection of strategies, maneuvers, and low-level control does not lead to contradicting selections of actions impacting their joint environment.
- If *ego(V)* and *ego(D)* are consistent, then beliefs of *ego(V||D)* are given by belief fusion, and actions are defined by parallel composition of their control components.

4 CONCLUSION

The formal semantic framework, sketched out above, demands a significant extension of known model classes. First, distribution-type beliefs about the state and intent of other interacting agents had to be rendered first-class members of the state space of the individual agents. It has been demonstrated in [8] and [10] that this integration of beliefs and evidence about other agents is a prerequisite for obtaining sound verdicts about the emergent joint behavior of ensembles of agents. Even the most expressive formal models lacking such reflexive state components, like hybrid automata [1] and their stochastic variants [14], are unable to properly encode rational interactive behavior and thus do necessarily yield behavioral verdicts that are either overly pessimistic or overly optimistic. Analyses of possible behavior to be expected from the system under design, like the PHA and HAZOP shown in Section 2 above, would consequently not be semantically well-founded if the underlying semantic model excluded and thereby ignored the impact of beliefs and evidence about other agents on individual and joint behavior. Belief about state and intent of other agents therefore had to become first-class citizens of the state set underlying the semantic foundation of the reference model. Having added belief about state and intent, such localized, individual belief of agents then naturally extends to the game arena within which the agents interact and that they exploit for computing their strategies: individual agents may have different perceptions of the options available to them and others and thus of the shape of the game arena in which they interact. Our model consequently not only permits inexact and possibly uncertain local copies of the game arena, but also their dynamic update based on observations, as observations may, for example, expose hitherto unexpected behavioral options that prompt belief revision concerning the structure of the game arena. Together, these elements induce a far more complex state space than reflected in the predominant hybrid-state models of cyber-physical systems, where the agent state "only" covers finite-dimensional vectors of real-valued and discrete state components [9].

Based on the rich semantic model exposed in Section 3, we can set out to realize multiple analysis methods. For example, the model enables preliminary HAZOP analysis to identify high risk scenarios, to derive time extended fault trees, to compute probabilities for minimal cut sets in such fault trees, and thus to derive causal models. We can analyze compliance with or violation of moral systems to the extent that their formal representation captures the intent. In principle, the model supports the derivation of abstractions supporting strategy synthesis with probabilistic guarantees of reaching a given set of goals, though actually performing such abstraction in a useful form is a non-trivial task currently left for future research. Likewise, the model supports the generation of test cases for analyzing the systems capability of maintaining safety in case of failures. Finally, we could subject models to attack scenarios testing the vulnerability of the system toward local or global cyber-attacks. In short, we provide a rigorous semantic foundation for such an endeavor. This foundation makes it possible eventually to develop and rigorously qualify pertinent tools, with the latter being subject of future research and development.

REFERENCES

[1] Rajeev Alur, Costas Courcoubetis, Thomas A. Henzinger, and Pei-Hsin Ho. 1993. Hybrid automata: An algorithmic approach to the specification and verification of hybrid systems. In *International Hybrid Systems Workshop*. Robert L. Grossman, Anil Nerode, Anders P. Ravn, and Hans Rischel (Eds.), Lecture Notes in Computer Science, Hybrid Systems, Vol. 736, Springer, Berlin, 209–229.

[2] Tamer Başar and Jan Olsder Geert. 1998. Dynamic noncooperative game theory. *Society for Industrial and Applied Mathematics*. https://doi.org/10.1137/1.9781611971132

[3] Klaus Bengler, Werner Damm, Andreas Luedtke, Rieger Jochem, Benedikt Austel, Bianca Biebl, Martin Fränzle, Willem Hagemann, Moritz Held, David Hess, Klas Ihme, Severin Kacianka, Alyssa J. Kerscher, Forrest Laine, Sebastian Lehnhoff, Alexander Pretschner, Astrid Rakow, Daniel Sonntag, Janos Sztipanovits, Maike Schwammberger, Mark Schweda, Anirudh Unni, and Eric Veith. 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. https://dl.acm.org/doi/abs/10.1145/3622880

[4] Dines Bjørner. 1979. The vienna development method (VDM). In *Proceedings of the Mathematical Studies of Information Processing*. E. K. Blum, M. Paul, S. Takasu (Eds.), Lecture Notes in Computer Science, vol 75, Springer, Berlin. DOI:https://doi.org/10.1007/3-540-09541-1_33

[5] A. Guy. 1998. *Boy: Cognitive function analysis*. Ablex, distributed by Greenwood Publishing Group, Westport, CT.

[6] Werner Damm, David Hess, Mark Schweda, Janos Sztipanovits, Klaus Bengler, Bianca Biebl, Martin Fränzle, Willem Hagemann, Moritz Held, Klas Ihme, Severin Kacianka, Alyssa J. Kerscher, Sebastian Lehnhoff, Andreas Luedtke, Alexander Pretschner, Astrid Rakow, Jochem Rieger, Daniel Sonntag, Maike Schwammberger, Benedikt Austel, Anirudh Unni, and Eric Veith. 2023. A reference architecture for human cyber physical systems- part I, in transactions on cyber-physical systems X(Y). *ACM (Association for Computing Machinery)*, New York, NY. https://dl.acm.org/doi/10.1145/3622879

[7] Lida David and Jan Maarten Schraagen. 2018. Analyzing communication dynamics at the transaction level: The case of air france flight 447. *Cognition, Technology & Work* 20, 4 (2018), 637–649.

[8] Martin Fränzle and Paul Kröger. 2018. The demon, the gambler, and the engineer - reconciling hybrid-system theory with metrology. *Symposium on Real-Time and Hybrid Systems* 11180 (2018), 165–185.

[9] Martin Fränzle, Mingshuai Chen, and Paul Kröger. 2019. In memory of oded maler: Automatic reachability analysis of hybrid-state automata. *ACM SIGLOG News* 6, 1 (2019), 19–39.

[10] Martin Fränzle and Paul Kröger. 2022. Bayesian hybrid automata: A formal model of justified belief in interacting hybrid systems subject to imprecise observation. *Leibniz Transactions on Embedded Systems* 8, 2 (2022), 05:1–05:27. DOI: https://doi.org/10.4230/LITES.8.2.5

[11] John C. Harsanyi. 1967. Games with incomplete information played by "Bayesian" players, I–III Part I. *The Basic Model. Management Science* 14, 3 (1967), 159–182.

[12] Duane Kritzinger. 2016. *Aircraft System Safety - Assessments for Initial Airworthiness Certification*. Elsevier, ISBN: 9780081008898.

[13] Nancy G. Leveson. 1995. *Safeware: System Safety and Computers*. ACM, NY. DOI: https://doi.org/10.1145/202709

[14] John Lygeros and Maria Prandini. 2010. Stochastic hybrid systems: A powerful framework for complex, large scale applications. *European Journal of Control* 16, 6 (2010), 583–594.

[15] Neville Moray. 1996. A taxonomy and theory of mental models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, 4 (1996), 164–168. DOI: https://doi.org/10.1177/154193129604000404

[16] Ian B. Rhodes and David G. Luenberger. 1969. Differential games with imperfect state information. *IEEE Transactions on Automatic Control* 14, 1 (1969), 29–38.

[17] S-18 Aircraft and Sys Dev and Safety Assessment Committee. 2010. *Guidelines for Development of Civil Aircraft and Systems Arp4754a*. SAE, (2010).

[18] Michael Stamatelatos, William Vesely, Joanne Dugan, Joseph Fragola, Jospeh Minarick III, and Kan Railsback. 2002. *Fault Tree Handbook With Aerospace Applications*. NASA.

[19] Neil Storey. 1996. *Safety Critical Computer Systems Addison-Wesley Longman Publishing Co.* Inc. Boston, MA, SBN:0201427877.

[20] Wolfgang Thomas. 1995. On the synthesis of strategies in infinite games. In *Proceedings of the Annual Symposium on Theoretical Aspects of Computer Science*. E. W. Mayr, C. Puech (Eds.), Lecture Notes in Computer Science. Springer, Berlin. DOI: https://doi.org/10.1007/3-540-59042-0_57

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