Grounding Embodied Multimodal Interaction

Towards Behaviourally Established Semantic Foundations for Human-Centred AI

Vasiliki Kondyli[1], Jakob Suchan[2] and Mehul Bhatt[1]

[1]*Örebro University, Sweden*

[2]*German Aerospace Center (DLR), Germany*

CoDesign Lab » Cognition. AI. Interaction. Design.
info@codesign-lab.org / https://codesign-lab.org

Abstract

We position recent and emerging research in *cognitive vision and perception* addressing three key questions: **(1)** What kind of relational abstraction mechanisms are needed to perform (explainable) grounded inference –e.g., question-answering, qualitative generalisation, hypothetical reasoning– relevant to embodied multimodal interaction? **(2)** How can such abstraction mechanisms be founded on behaviourally established cognitive human-factors emanating from naturalistic empirical observation? and **(3)** How to articulate behaviourally established abstraction mechanisms as formal declarative models suited for grounded knowledge representation and reasoning (KR) as part of large-scale hybrid AI and computational cognitive systems.

We contextualise (1–3) in the backdrop of recent results at the interface of AI/KR, and Spatial Cognition and Computation. Our main purpose is to emphasise the importance of behavioural research based foundations for next-generation, human-centred AI, e.g., as relevant to applications in Autonomous Vehicles, Social and Industrial Robots, and Visuo-Auditory Media.

Keywords

Multimodal Interaction, Commonsense Reasoning, Declarative Spatial Reasoning, Declarative AI, Explainable AI, Cognitive Human-Factors, Cognitive Systems

1. Motivation

Multimodality in interaction is an inherent aspect of human activity, be it in social, professional, or everyday mundane contexts. Next-generation AI technologies, aiming for compliance with human-centred ethical and legal requirements, performance benchmarks, and inclusive usability expectations will require an inherent foundational capacity to analyse –e.g., understand, explain, anticipate– everyday interactional multimodality in naturalistic settings involving technology mediated collaborative assistance of humans. Amongst other things, this necessitates that the foundational building blocks of such next-generation systems be semantically aligned with the descriptive complexity of human task conceptualisation and performance expectations.

Declaratively Mediated Multimodality. The significance of "*grounding*" in semiotic construction, e.g., enabling high-level meaning-making, has been long-established in Artificial

KR4HI 2022: Knowledge Representation for Hybrid Intelligence, June 14, 2022, Amsterdamm, The Netherlands

vasiliki.kondyli@oru.se (V. Kondyli); jakob.suchan@dlr.de (J. Suchan); mehul.bhatt@oru.se (M. Bhatt)

Intelligence and related disciplines. Our research addresses the theoretical, methodological, and applied understanding of "grounded representation" mediated multimodal sensemaking of human behaviour at the interface of language, logic, and cognition [1]. Here, declaratively mediated grounded inference for collaborative autonomy through systematic neurosymbolic mechanisms integrating knowledge representation and reasoning with visual computing is of special significance. Intended functional purposes encompass diverse operative needs such as explainable multimodal commonsense understanding, multimodal generation/synthesis for communication and summarisation, multimodal interpretation guided decision-support, multimodal behaviour adaptation & autonomy, and multimodal analytical visualisation. It is also necessary that methods and tools developed towards realising such operational needs are designed tot be domain agnostic, and that they cater to both online/real-time as well as post-hoc operation in diverse application scenarios (e.g., refer [2] for the case of online neurosymbolic abduction applied to the domain of autonomous driving).

Behavioural Foundations: Cognitive Human-Factors Informing KR. By analysing real-world everyday scenarios that involve interactions between humans as well as between humans and their surrounding environment, we extract and categorize cognitive aspects of human multimodal communication and explore the nature of human interactions under different circumstances (e.g. education, collaborative tasks, driving, social communication, industrial work), as well as the effect of external factors (e.g. environmental complexity, event complexity) on the interactions [3] (Fig. 1). Our research focus is on cognitive human-factors, primarily pertaining to visual attention, and multimodal interpersonal communication that can be studied in real-world scenes as well as through behavioural naturalistic studies. Through a systematic study of such cognitive human-factors, e.g., involving conceptual and as well as empirical analyses, we explicate means and mechanisms that are critical in human multimodal interaction; these in turn constitute the basis of formal model development informing the design and development human-centred AI systems.

In this position statement, we aim to present the confluence of the aforestated computational and behavioural aspects of our research. Our aim is to conceptually highlight recent and emerging work, while pointing out interested readers to relevant publications –e.g., particularly [2, 4, 5, 6, 7]– where furthher (KR-centric) details may be consulted.

2. Cognitive Human-Factors Concerning Multimodality

Multimodal interpersonal communication refers to exchange of signals involving speech, gestures, facial expressions, body posture and more. Multiple times people use a combination of these signals to achieve a higher level of social interaction and mutual understanding of a situation (e.g. joint or shared attention) [8]. The multimodality of an interaction refers to a person's way of communicating by using more than one modality at the same time as a signal, or from a perceptual approach, it refers to more than one modality being received based on the receiver's perception of the signal [9]. Moreover, multimodal interactions are also involved in everyday activities where people proceed to a sequence of actions in which interact with other agents or objects (e.g. cooking, assembling).

From a behavioural viewpoint, we investigate the nature of multimodal interaction in diverse contexts as applicable to a range of application areas of interest, e.g., everyday driving [10, 11], media studies [12, 13], social interaction and activity performance [14, 15] as applicable to autonomous systems contexts as follows (Fig. 1):

- **Autonomous Driving**. Interactions between street stakeholders such as drivers, pedestrians, cyclists, etc. are mostly characterised by non-verbal communication, involving a range of multimodal signals such as gestures, head movements, eye contact, that are either based on traffic rules or socially and culturally developed to resolve traffic ambiguities. We are interested in the combination of multimodal interactions used to establish joint attention between the roadside users, especially during safety critical situations, and the manner in which this knowledge of efficient coordination can be transfered to human and autonomous systems communication (Fig. 1a).
- **Visuo-auditory Media**. In media communication, even the most "verbal" communication as media news makes use of the affordances provided by various non-verbal modes of communication, such as prosody, intonation, gestures or facial expressions. The analysis of visual and auditory stimuli in combination with visual attention, and reactions by the audience lead to a better understanding of the effect that even humble changes in gaze direction or in body posture can have in audience's experience (Fig. 1b).
- **Social and Industrial Robotics**. Collaborative tasks among humans, or humans and social robots, intelligent systems, or robotic arms, require a high-level of awareness about the status, the actions and the interactions of the other part of the collaboration that is usually expressed through multimodal actions tightly related to the nature of the task performed. Head movement, gaze, and body posture, gaze are indicative in these interactions, while the special subcategories of deictic and iconic gestures are present in assembling everyday tasks such as cooking, organising, transferring, etc (Fig. 1c - 1d).

By analysing interaction events from various settings, and using knowledge of human perception and cognition during such interactions, we can explicate aspects of communication that are frequently connected to a successful or a failed interaction. A successful interaction, e.g., in settings such as in Fig. 2, is frequently characterised by mutual social attention, and it can be achieved with multiple *modes* and *means* of delivering intentions (e.g. explicit, implicit) through a combination of (visuo-auditory) modalities (e.g. gestures, head movements, speech and intonation). To study the range of possible interactions and their effect on human behaviour, we conducted empirical studies focusing on evidence based qualitative analysis of embodied multimodal interactions in naturalistic situations in two contexts: (1) embodied decision-making in everyday driving, including a number of everyday interaction scenarios in real-world scenes as well as in the virtual environment (e.g., crossing a street, overtaking, avoiding an object, etc.) [16]; (2) visual attention in the moving image, where we analyse the changes in visual attention along various movie clips (e.g. audience attention follows gaze swifts by actors, directing attention by gestures). In these studies we collect and analyse multimodal data as seen in Table 1. That led to an extended dataset of dynamic naturalistic stimuli accompanied by empirical evidence for evaluating human performance under different interaction events, as well as in different levels of visuospatial and event complexity.

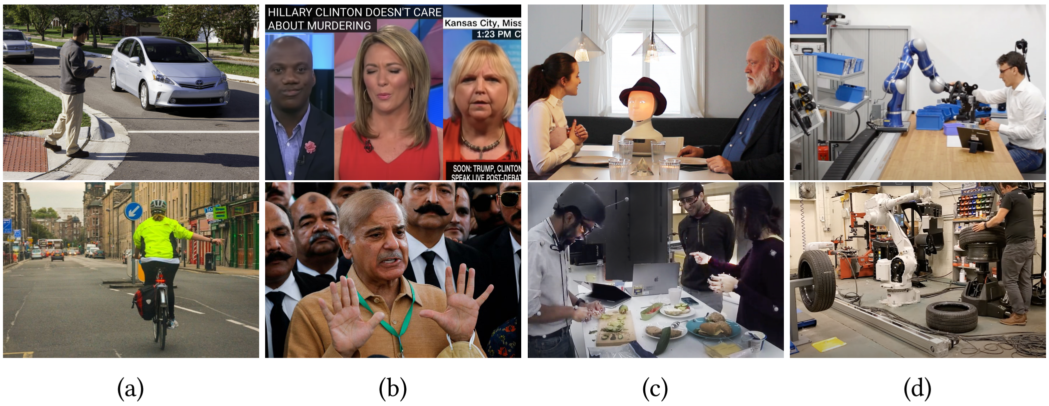

HILLARY CLINTON DOESN'T CARE ABOUT MURDERING
Kansas City, Miss
1:23 PM C
SOON: TRUMP, CLINTON SPEAK LIVE POST-DEBAT
(a)
(b)
(c)
(d)

Figure 1: Multimodal communication in various contexts: **(a)** A pedestrian establishes joint attention with a driver, and a cyclist's gesture indicates intentions to turn following traffic rules; **(b)** Facial expressions accompany speech during news media discussions or public talks; **(c)** Eye contact and deictic gestures promoting joint attention under social (robotic) collaborative tasks; **(d)** Industrial collaborative tasks with a robotic arm.

Behavioural Analysis –	**Human Perception Metrics**
Performance Evaluation (task specific)	Search time - Accuracy - Detection rate - Detection time - Reaction time Steering - Breaking - Accelerating - Time of completion
Physiological Measurements	Eye-tracking: Latency of first saccade - Number of fixations - Number of fixations (on targets) - Number of fixations (on distractors) - Duration of Fixations - Scanpath ration - Final saccade length - Pupilometry
Behavioural Metrics	Head Movements - Think aloud - Sketch map - Orientation tasks - Changes in Speed

Table 1
Summary of metrics used for behavioural evaluation through empirical data.

3. Grounded Representation Mediated Multimodality

Informed by the behavioural research into cognitive human-factors pertaining multimodal interaction, we develop formal (declarative) methods supporting grounded relational categorisations –pertaining to space, motion, events, and actions– linking linguistic expressions with non-linguistic, especially quantitative perceptual data pertaining, for instance, dynamic spatio-temporal phenomena in embodied interaction contexts. Multimodal interpretation in our work is broadly construed in the context of diverse forms of imagery, e.g., encompassing perceptual and communicative data sources such as image, video, language, audition, text, eye-tracking, neurophysiological markers in behavioural / clinical settings (Table 1). Example applications in focus include autonomous vehicles, social & industrial robotics, creative design technologies, and clinical diagnostic intervention tools. For the purposes of this position statement, we focus on two select examples: joint attention in everyday driving, and everyday-activity-related perceptual grounding in a robotics setting.

<table>
<tr><th colspan="3">Grounded Representation</th></tr>
<tr><th colspan="3">Interactions and Communication Tools</th></tr>
<tr><td>Practical Action</td><td>Object / Environment Interactions - Auditory cues - Motion Paths</td><td>enters(P,Q), crossing(P,Q), passing_behind(P,Q), hides_behind(P,Q), approaching(P,Q), opening(P,Q), removing(P,Q), holding(P,Q), touching(P,Q), ...</td></tr>
<tr><td>Explicit Interaction</td><td>Eye Contact - Facial Expressions - Gestures - Speech - Nodding</td><td>joint_attention(P,Q), monitoring_attention(P,Q), gesture(P, Gesture), hand_sign(P, Sign), auditory_cue(Source, Cue), ...</td></tr>
<tr><td>Implicit Interaction</td><td>Body Posture / Positioning - Head Movement - Gaze - Intonation - Behavioural changes</td><td>pose(P,Pose), turn_head(P, Direction), speed_up(P), maintain_steady_speed(P), slow_down(P), detect(P,Q), track(P,Q), ...</td></tr>
<tr><th colspan="3">Facts / Beliefs (Fluents)</th></tr>
<tr><td>Scene Properties</td><td colspan="2">visibility: hidden(P), partially_hidden(P), occluded_by(P, Q), , ...; attention: looking_at(P, Q), attentive(P), ...; location: on(P, Q), in(P, Q), next_to(P, Q), ...</td></tr>
<tr><th colspan="3">Scene Elements</th></tr>
<tr><td>Types (Taxonomy)</td><td colspan="2"><table>
<tr><td rowspan="3">object</td><td rowspan="2">dynamic</td><td colspan="2">person, animal, ...</td></tr>
<tr><td>vehicle</td><td>car, truck, motorcycle, bicycle, ...</td></tr>
<tr><td>static</td><td colspan="2">traffic light, barrier, obstacle, ...</td></tr>
</table>
<table>
<tr><td>region</td><td>road, sidewalk, lane, intersection ...</td></tr>
</table></td></tr>
<tr><td>Structure & Properties</td><td colspan="2">people: body-parts (hands, face, ...), body pose, facing direction, gaze direction, ...
objects: orientation, parts, ...</td></tr>
<tr><th colspan="3">Spatio-Temporal Characterisation</th></tr>
<tr><td>Domains</td><td colspan="2">Mereotopology, Incidence, Orientation, Distance, Size, Motion, ...</td></tr>
<tr><td>Relations</td><td colspan="2">topology / position: inside, outside, overlapping, connected, left, right, in front, behind, on top, touching; direction: facing towards, facing away, same direction, opposite direction; moving: towards, away, parallel; ...</td></tr>
<tr><td>Entities</td><td colspan="2">bounding boxes, polygons, line-segments, points, oriented-points, motion trajectories, time-points, time intervals, ...</td></tr>
</table>

Table 2
Deep Semantic Structure of Multimodal Interactions

3.1. Grounding Multimodal Interactions: An Ontological Characterisation

From the analysis of real-world scenarios, as well as real-world based behavioural studies using perceptual metrics to assess human behaviour (Table 1), we extract common features of interactions, that can be categorised in a high-level, based on the type of (inter)action, the modalities as communication tools, while the analysis of the scene leads to a definition of facts and beliefs that describe the observations concerning the sequence of events. For example, observing eye contact between a pedestrian and a driver, as well as the fact that the traffic light is visible to both parts and green for the pedestrian, while the pedestrian is positioned close to the street, lead to an informed assumption that the pedestrian feels safe and intends to cross the road. These high-level interpretations of events can be further described with low-level elements, involving objects and areas of the scene, their properties as well as the changes in their spatio-temporal relations as the scenario evolves. We provide a structure of deep semantics for multimodal interactions that is common to various domains of application, and which includes high and low-level representations based on the following categories (Table 2):

» **Interactions and Communication Tools**. Interactions are characterised based on the relational spatio-temporal structure underlying the respective interaction and the effects on the facts and beliefs about the world they are performed in. Practical actions (e.g. (re)direction of a path, pushing/pulling an object), describe the interactions between a person and the environment during an everyday task. Communicative interactions are classified based on the mode of deliverance of the message, as explicit or implicit interactions. Explicit interactions involve a range of modalities such as facial expressions or gestures, e.g. a cyclist's extension of one hand on the side, is a gesture that conveys his intention to turn in the upcoming intersection (Fig. 1a). Implicit interactions involve a set of modalities as communication tools that require lower effort, such as gaze, body posture or head movements, e.g. a gaze shift the someone's mobile phone towards the street traffic might communicate his intention to cross (Fig. 1a).

» **Facts – Beliefs**. The observed and inferred environmental properties and characteristics of the entities in a scene, such as the visibility of objects/agents, their locations and facing directions, etc., are considered facts and beliefs that describe the state of the world. A combination of facts and events observed over a longer time interval may lead to hypothesis about ongoing interactions, agent's intentions, or the anticipation of near future events. In particular, analysing the sequence of interactions, as well as the properties of a scene, and the resulting changes in the belief state lead to hypothesis about the events occurring, or being in progress. For instance, observing a pedestrian located next to a zebra crossing while he is looking towards the road traffic indicates an intention of crossing in the near future.

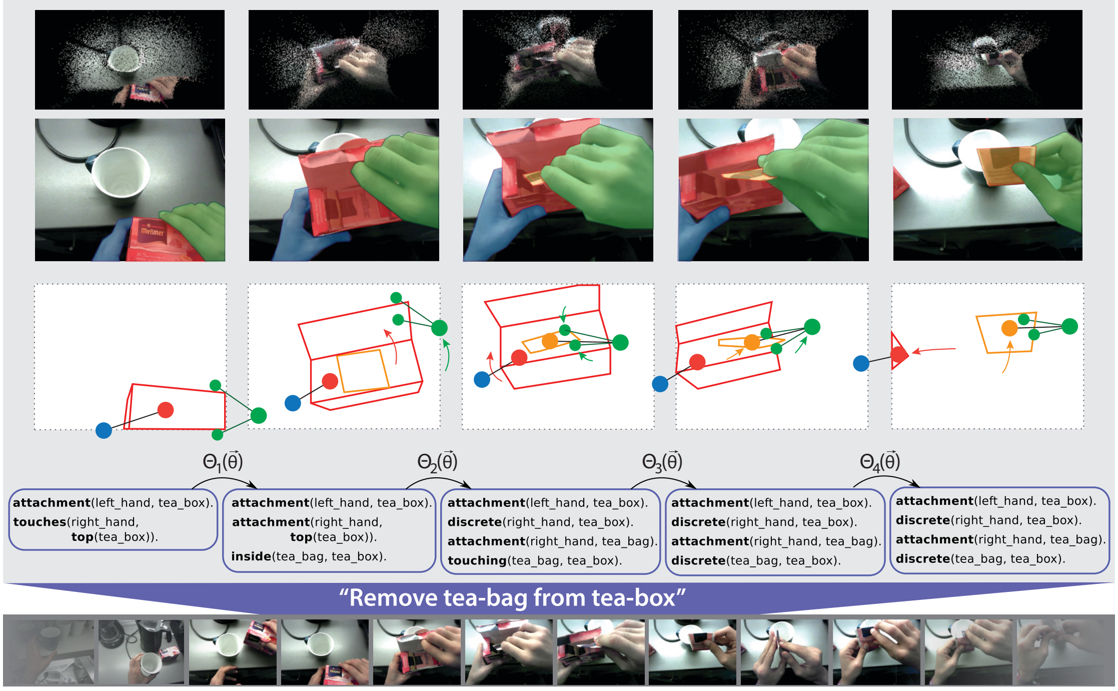
» **Scene Elements**. The distinct, domain specific elements of the physical world obtained from high-level sensing and processing and describing the scene, e.g. traffic lights, zebra crossings for a driving scenario, or right hand, tea box, table for an everyday action scenario. These elements are categorised based on their type, structure and properties, and are geometrically represented as low-level entities (e.g. bounding boxes) that are involved in spatio-temporal relationships, and that constitute the underline representations to describe interactions.

» **Spatio-Temporal Characterisation**. Referring to commonsense relations for the abstraction of space, motion, and (inter)action. These involve primitive spatio-temporal entities and relations holding amongst them, with respect to position, orientation, direction of movement, etc., during a time interval. An adequate commonsense spatio-temporal characterisation can connect with low-level quantitative data, and also help to ground symbolic descriptions of actions and objects to be queried, reasoned about, or even manipulated in the real world.

EXAMPLES I–II. In two scenarios from two different domains: (**I**) everyday actions, and (**II**) driving, we show how high-level interactions between humans as well as between human and objects can be represented based on the deep semantic structure introduced.

I. Perceptual Grounding of Everyday Activities. Analysing an everyday scenario of human activity such as "making a cup of tea" (Fig. 2) from the perspective of the person, may be interpreted as a sequence of interactions represented as high-level steps such as:

> opening the tea-box, removing a tea-bag from the box and putting the tea-bag into a cup filled with water while holding the cup.



attachment(left_hand, tea_box).
touches(right_hand,
top(tea_box)).

attachment(left_hand, tea_box).
attachment(right_hand,
top(tea_box)).
inside(tea_bag, tea_box).

attachment(left_hand, tea_box).
discrete(right_hand, tea_box).
attachment(right_hand, tea_bag).
touching(tea_bag, tea_box).

attachment(left_hand, tea_box).
discrete(right_hand, tea_box).
attachment(right_hand, tea_bag).
discrete(tea_bag, tea_box).

attachment(left_hand, tea_box).
discrete(right_hand, tea_box).
attachment(right_hand, tea_bag).
discrete(tea_bag, tea_box).

Figure 2: Analysing interactions with an object in an everyday activity – "making a cup of tea" (egocentric view from a head-mounted RGB-D capture device).

Each action can be described with high-level spatial and temporal relationships between the person and the involved objects. For example, one may identify relationships of contact and containment that hold across specific time-intervals, such as:

> the left hand is attached to the tea box, the right hand is touching the tea box at the top, then the right hand is touching a tea bag, while the tea bag is inside the tea box. After that, the right hand is moving away from the tea box together with the tea bag.

The parametrised manipulation or control actions[1] $(\Theta_1(\theta), \ldots \Theta_n(\theta))$ effectuate state transitions, which may be qualitatively modelled as changes in topological relationships amongst involved domain entities.

II. Joint Attention in Everyday Driving. Of special interest are multimodal interactions between the different street stakeholders (e.g. drivers, cyclist, pedestrians) in safety critical situations during everyday driving scenarios, and the level of (mutual) social attention achieved between the parts during the course of interaction. Using the semantic structures of Table 2, in two successive instances, we describe the sequence of multimodal interactions between the involved parts (Fig. 3):

[1] Control actions are formally defined based on their preconditions and effects and are used to represent spatio-temporal interactions in the scene. For a sample technical elaboration, consult [17]. A more broader discussion on the intended/possible semantics of integrated “*reasoning about space, actions, and change*” is available in [18], where formal semantics realised in a *situation calculus* setting is available in [19].

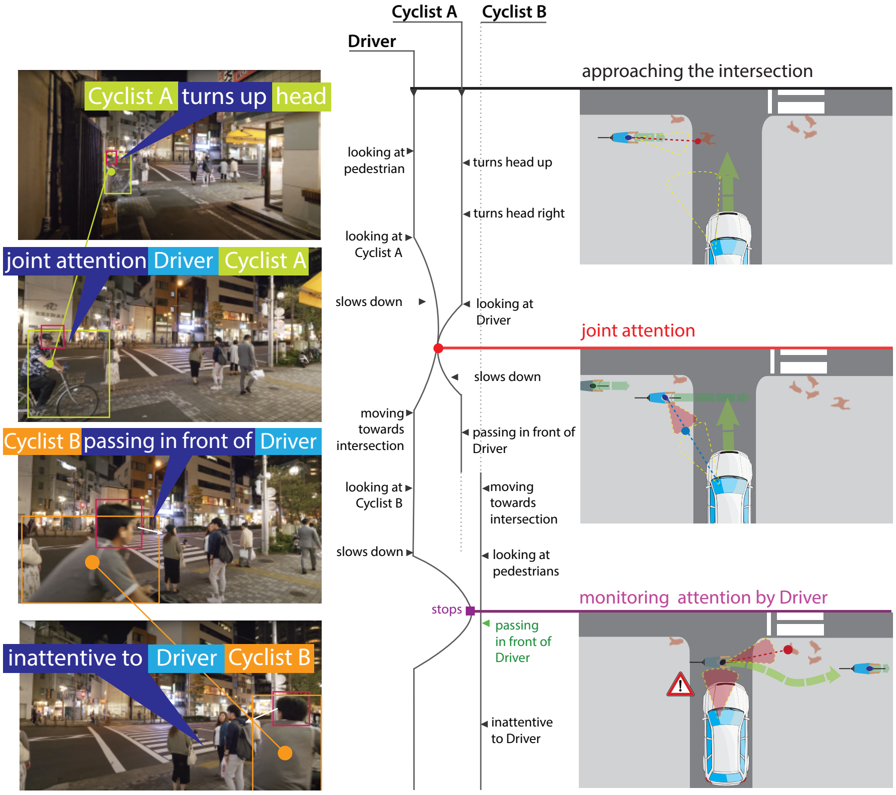

Cyclist A turns up head
joint attention Driver Cyclist A
Cyclist B passing in front of Driver
inattentive to Driver Cyclist B
Cyclist A
Cyclist B
Driver
approaching the intersection
looking at pedestrian
turns head up
turns head right
looking at Cyclist A
slows down
looking at Driver
joint attention
slows down
moving towards intersection
passing in front of Driver
looking at Cyclist B
moving towards intersection
slows down
looking at pedestrians
monitoring attention by Driver
stops
passing in front of Driver
inattentive to Driver

Figure 3: Analysing the sequence of two interactions: **Interaction 1** – between Driver and Cyclist A, and **Interaction 2** – between the Driver and Cyclist B. This instance involves a successful and a failed case of establishing joint attention between the two parts of interaction.

> **Interaction 1:** A Driver is located on a one-way road and is moving towards the intersection, while Cyclist A is located on the sidewalk on the left of the Driver and is moving towards the same intersection. Boths part are approaching the intersection. The Driver looks at Cyclist A. When Cyclist A is at a distance of 10 meters from the intersection, Cyclist A performs a head turn towards right and he is looking at the Driver. As a result, they establish joint attention. Then, the Driver and Cyclist A slow down, and then Cyclist A passes in front of Driver. After Cyclist A crosses the intersection, the Driver moves towards the intersection, until the time when Cyclist A is not visible in the scene anymore.

The two parts of the interaction at this instance, Driver and Cyclist A, are approaching the same intersection from two different directions. Because we examine the instance from the perspective of the Driver, we observe Cyclist A approaching the intersection from the left side of the scene. The interaction between the two parts starts with monitoring attention from the Driver, who detects and tracks Cyclist A (Gaze), meaning that at that stage only the Driver is aware of the situation and can anticipate the upcoming events close to the intersection. Then, practical actions are getting involved in the interaction, when the Driver slows down as a reaction to the monitoring attention he establishes with Cyclist A, and later implicit interactions are introduced when Cyclist A turns his head towards the Driver (Head Movement) and looks at the Driver (Gaze). After these multimodal signals, the two parts engage in eye contact, and so the interaction evolves into an explicit interaction of joint attention. At the stage of joint attention, is expected that the two parts of the interaction have achieved a mutual understanding of the situation, and that they can proceed to a safe resolution of the episode. As we can confirm from the observations after this interaction, the two parts are crossing the intersection one after the other, that indicates an efficient communication.

Similarly, the second interaction can be represented as follows:

> **Interaction 2:** The Driver is looking at Cyclist B, while Cyclist B is looking towards the pedestrians. The Driver slows down, while Cyclist B maintains steady speed. The Driver stops and Cyclist B is passing in front of the Driver in close distance.

The Driver detects and tracks Cyclist B along the way, while Cyclist B appears inattentive towards the Driver. Cyclist B does not gaze towards the car, but instead he tracks the movements of pedestrians close to the intersection. In this case, the Driver establishes monitoring one-sided attention with Cyclist B. As there is no change in the behaviour of Cyclist B during the interaction, the social attention remains in the level of monitoring attention from the perspective of the Driver because there is no evidence of any mutual understanding of the situation by Cyclist B during the time interval we examine. As a result, the Driver slows down, and eventually stops to avoid an accident with Cyclist B. In this case we observe how an one-sided attention, and a failed communication between the parts can lead to a "close to accident" event.

By examining the two interactions together we observe that there is a narrow margin in time between the two incidents and for this reason the Driver had to switch his attention quickly from Cyclist A to Cyclist B. Consequently, failure in interactions, and respectively failure to a mutual understanding of the situation by all the parts involved can introduce major load to one part, the Driver in our case. This is a common example of how environmental factors, including visuospatial complexity (e.g. narrow street, low visibility from the sidewalk) as well as events complexity (e.g. multiple events in short time) can affect the interactions and lead to safety critical situations.

3.2. Deep Semantic Inference

The development of domain-independent computational models of perceptual sensemaking — e.g., encompassing visuospatial Q/A, learning, abduction— with multimodal human behavioural stimuli such as RGB(D), video, audio, eye-tracking requires the representational and inferential mediation of commonsense and spatio-linguistically rooted abstractions of space, motion, actions, events and interaction (Table 2). We characterise deep (visuospatial) semantics as:

- the existence of declarative models pertaining to space, space-time, motion, actions & events, spatio-linguistic conceptual knowledge and their corresponding formalisation supporting (domain-neutral) commonsense cognitive reasoning capabilities with quantitatively sensed dynamic visual imagery. Here, it is of the essence that an expressive ontology consisting of, for instance, space, time, space-time motion primitives as first-class objects is accessible within the (declarative) programming paradigm under consideration, and that operational (reasoning) capabilities such as visuospatial question-answering, spatio-temporal learning, non-monotonic visuospatial abduction be directly supported.

We particularly emphasise the abilities to **abstract, learn, and reason** with cognitively rooted structured characterisations of commonsense knowledge about **space and motion**. Formal semantics and computational models of deep semantics manifest themselves in declarative AI settings such as Constraint Logic Programming (CLP) [20], Inductive Logic Programming (ILP) [21], and Answer Set Programming (ASP) [22]. Present focus has been on visuospatial question-answering, abduction, and relational learning:

I. Visuospatial Question-Answering [6, 23]. Focus is on a computational framework for semantic-question answering with video and eye-tracking data founded in constraint logic programming; we also demonstrate an application in cognitive film & media studies, where human perception of films vis-a-via cinematographic devices is of interest.

II. Visuospatial Abduction [4, 7, 24]. Focus is on a hybrid architecture for systematically computing robust visual explanation(s) encompassing hypothesis formation, belief revision, and default reasoning with video data (for active vision for autonomous driving, as well as for offline processing). The architecture supports visual abduction with *space-time histories* as native entities, and founded in (functional) answer set programming based spatial reasoning.

III. Relational Visuospatial Learning [5, 25]. Focus is on a general framework and pipeline for: relational spatio-temporal (inductive) learning with an elaborate ontology supporting a range of space-time features; and generating semantic, (declaratively) explainable interpretation models in a neurosymbolic pipeline demonstrated for the case of analysing visuospatial symmetry in visual art.

From a foundational viewpoint, a deep semantic (grounded) multimodal inference entails inherent support for tackling a range of challenges concerning *epistemological* and *phenomenological* aspects relevant to a wide range of *dynamic spatial systems* [19, 26, 18]:

- **interpolation and projection** of missing information, e.g., what could be hypothesised about missing information (e.g., moments of occlusion); how can this hypothesis support planning an immediate next step?
- object **identity maintenance** at a semantic level, e.g., in the presence of occlusions, missing and noisy quantitative data, error in detection and tracking,
- ability to make **default assumptions**, e.g., pertaining to persistence of objects and/or object attributes,
- maintaining **consistent beliefs** respecting (domain-neutral) commonsense criteria, e.g., related to compositionality & indirect effects, space-time continuity, positional changes resulting from motion,
- inferring / computing **counterfactuals**, in a manner akin to human cognitive ability to perform mental simulation for purposes of introspection, performing "what-if" reasoning tasks etc.

Addressing such challenges —be it realtime or post-hoc— in view of human-centred AI concerns pertaining to representations rooted to natural language, explainability, ethics and regulation requires a systematic (neurosymbolic) integration of **Semantics and Vision**, i.e., robust commonsense representation & inference about *spacetime dynamics* on the one hand, and powerful low-level visual computing capabilities, e.g., pertaining to object and other human feature detection and tracking.

4. Conclusion

Semantically grounded reasoning (e.g., with sensor data pertaining to multimodal human behaviour [27]) has been long recognised to be a crucial requirement to achieve computational cognition. Yet, its significance must now be reiterated, re-asserted even, in view of recent advances in neural machine learning, and its status quo vis-a-vis explainability requirements from the viewpoint of human-centred AI. We suggest that KR-research has always concerned itself with the "hard" question of *semantics*, entailing explainability amongst other things, and that KR research and its role and contribution towards large-scale hybrid intelligence is of even greater significance now than ever before given the tremendous opportunities afforded by methods such as deep learning. In this position statement, we have attempted to summarise our ongoing work towards establishing a human-centric foundation and roadmap for the development of neurosymbolically grounded inference about embodied multimodal interaction as relevant to a range of application contexts. For key technical details and to obtain a summary of open directions, we direct interested readers to [2, 4, 5, 6, 7].

References

[1] M. Bhatt, J. Suchan, Cognitive vision and perception, in: G. D. Giacomo, A. Catalá, B. Dilkina, M. Milano, S. Barro, A. Bugarín, J. Lang (Eds.), ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020), volume 325 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2020, pp. 2881–2882. URL: https://doi.org/10.3233/FAIA200434. doi:10.3233/FAIA200434.

[2] J. Suchan, M. Bhatt, S. Varadarajan, Commonsense visual sensemaking for autonomous driving - on generalised neurosymbolic online abduction integrating vision and semantics, Artif. Intell. 299 (2021) 103522. URL: https://doi.org/10.1016/j.artint.2021.103522. doi:10.1016/j.artint.2021.103522.

[3] V. Kondyli, M. Bhatt, J. Suchan, Towards a human-centred cognitive model of visuospatial complexity in everyday driving, in: CEUR Workshop Proceedings :, volume 2655 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2655/paper20.pdf.

[4] J. Suchan, M. Bhatt, P. A. Walega, C. Schultz, Visual explanation by high-level abduction: On answer-set programming driven reasoning about moving objects, in: S. A. McIlraith, K. Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 1965–1972. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17303.

[5] J. Suchan, M. Bhatt, S. Vardarajan, S. A. Amirshahi, S. Yu, Semantic Analysis of (Reflectional) Visual Symmetry: A Human-Centred Computational Model for Declarative Explainability, Advances in Cognitive Systems 6 (2018) 65–84. URL: http://www.cogsys.org/journal.

[6] J. Suchan, M. Bhatt, Semantic question-answering with video and eye-tracking data: AI foundations for human visual perception driven cognitive film studies, in: S. Kambhampati (Ed.), Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, IJCAI/AAAI Press, 2016, pp. 2633–2639. URL: http://www.ijcai.org/Abstract/16/374.

[7] J. Suchan, M. Bhatt, S. Varadarajan, Out of sight but not out of mind: An answer set programming based online abduction framework for visual sensemaking in autonomous driving, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 1879–1885. doi:10.24963/ijcai.2019/260.

[8] S. Oviatt, R. Coulston, R. Lunstord, When do we interact multimodally? cognitive load and multimodal communication patterns, ICMI, Pennsylvania, USA, 2004.

[9] C. Smith, C. Evans, A new heuristic for capturing the complexity of multimodal signals, Behavioral Ecology and Sociobiology 67 (2013).

[10] R. Florian, Natural Multimodal Interaction in the Car - Generating Design Support for Speech, Gesture, and Gaze Interaction while Driving, Ph.D. thesis, Bamberg, 2021. URL: https://fis.uni-bamberg.de/handle/uniba/51826. doi:10.20378/irb-51826.

[11] A. Rasouli, I. Kotseruba, K. J. Tsotsos, Are they going to cross? a benchmark dataset

and baseline for pedestrian crosswalk behavior, in: IEEE Inter Confon Computer Vision Workshops, 2017, pp. 206–213.

[12] F. Steen, M. B. Turner, Multimodal Construction Grammar, Stanford, CA: CSLI Publications, 2012.

[13] J. Bateman, K.-H. Schmidt, Multimodal Film Analysis: how films mean, London: Routledge, 2011.

[14] G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, E. André, Exploring a model of gaze for grounding in multimodal hri, in: Proceedings of the 16th International Conference on Multimodal Interaction, ICMI '14, Association for Computing Machinery, New York, NY, USA, 2014, pp. 247–254.

[15] E. André, Socially interactive artificial intelligence: Past, present and future, in: Proceedings of the 2021 International Conference on Multimodal Interaction, ICMI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 4.

[16] V. Kondyli, M. Bhatt, Multimodality on the road : Towards evidence-based cognitive modelling of everyday roadside human interactions, in: Advances in Transdisciplinary Engineering :, volume 11 of *Advances in Transdisciplinary Engineering*, IOS Press, 2020, pp. 131–142. doi:10.3233/ATDE200018.

[17] J. Suchan, M. Bhatt, Commonsense Scene Semantics for Cognitive Robotics: Towards Grounding Embodied Visuo-Locomotive Interactions, in: ICCV 2017 Workshop: Vision in Practice on Autonomous Robots (ViPAR), International Conference on Computer Vision (ICCV), 2017.

[18] M. Bhatt, Reasoning about Space, Actions and Change: A Paradigm for Applications of Spatial Reasoning, in: Qualitative Spatial Representation and Reasoning: Trends and Future Directions, IGI Global, USA, 2012.

[19] M. Bhatt, S. W. Loke, Modelling dynamic spatial systems in the situation calculus, Spatial Cognition & Computation 8 (2008) 86–130. URL: https://doi.org/10.1080/13875860801926884. doi:10.1080/13875860801926884.

[20] J. Jaffar, M. J. Maher, Constraint logic programming: A survey, The journal of logic programming 19 (1994) 503–581.

[21] S. Muggleton, L. D. Raedt, Inductive logic programming: Theory and methods, Journal of Logic Programming 19 (1994) 629–679.

[22] G. Brewka, T. Eiter, M. Truszczyński, Answer set programming at a glance, Commun. ACM 54 (2011) 92–103. doi:10.1145/2043174.2043195.

[23] J. Suchan, M. Bhatt, The geometry of a scene: On deep semantics for visual perception driven cognitive film, studies, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016, IEEE Computer Society, 2016, pp. 1–9. URL: https://doi.org/10.1109/WACV.2016.7477712. doi:10.1109/WACV.2016.7477712.

[24] P. A. Walega, M. Bhatt, C. P. L. Schultz, ASPMT(QS): non-monotonic spatial reasoning with answer set programming modulo theories, in: F. Calimeri, G. Ianni, M. Truszczynski (Eds.), Logic Programming and Nonmonotonic Reasoning - 13th International Conference, LPNMR 2015, Lexington, KY, USA, September 27-30, 2015. Proceedings, volume 9345 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 488–501. URL: https://doi.org/10.1007/978-3-319-23264-5_41. doi:10.1007/978-3-319-23264-5_41.

[25] J. Suchan, M. Bhatt, C. P. L. Schultz, Deeply semantic inductive spatio-temporal learning, in: J. Cussens, A. Russo (Eds.), Proceedings of the 26th International Conference on Inductive Logic Programming (Short papers), London, UK, 2016, volume 1865, CEUR-WS.org, 2016, pp. 73–80.

[26] M. Bhatt, H. W. Guesgen, S. Wölfl, S. M. Hazarika, Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions, Spatial Cognition & Computation 11 (2011) 1–14. URL: https://doi.org/10.1080/13875868.2010.548568. doi:10.1080/13875868.2010.548568.

[27] M. Bhatt, K. Kersting, Semantic interpretation of multi-modal human-behaviour data - making sense of events, activities, processes, KI / Artificial Intelligence 31 (2017) 317–320.

[28] M. Bhatt, J. Suchan, S. Vardarajan, Deep semantics for explainable visuospatial intelligence : Perspectives on integrating commonsense spatial abstractions and low-level neural features, in: Proceedings of the 2019 International Workshop on Neural-Symbolic Learning and Reasoning : Annual workshop of the Neural-Symbolic Learning and Reasoning Association, 2019. URL: https://www.researchgate.net/publication/333480472_Deep_Semantics_for_Explainable_Visuospatial_Intelligence_Perspectives_on_Integrating_Commonsense_Spatial_Abstractions_and_Low-Level_Neural_Features.

[29] J. Suchan, M. Bhatt, Deep Semantic Abstractions of Everyday Human Activities: On Commonsense Representations of Human Interactions, in: ROBOT 2017: Third Iberian Robotics Conference, Advances in Intelligent Systems and Computing 693, 2017.

[30] M. Eppe, M. Bhatt, Approximate postdictive reasoning with answer set programming, J. Appl. Log. 13 (2015) 676–719. URL: https://doi.org/10.1016/j.jal.2015.08.002. doi:10.1016/j.jal.2015.08.002.

[31] M. Eppe, M. Bhatt, F. Dylla, Approximate epistemic planning with postdiction as answer-set programming, in: P. Cabalar, T. C. Son (Eds.), Logic Programming and Nonmonotonic Reasoning, 12th International Conference, LPNMR 2013, Corunna, Spain, September 15-19, 2013. Proceedings, volume 8148 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 290–303. URL: https://doi.org/10.1007/978-3-642-40564-8_29. doi:10.1007/978-3-642-40564-8_29.

[32] M. Eppe, M. Bhatt, A history based approximate epistemic action theory for efficient postdictive reasoning, J. Appl. Log. 13 (2015) 720–769. URL: https://doi.org/10.1016/j.jal.2015.08.001. doi:10.1016/j.jal.2015.08.001.

[33] M. Spranger, J. Suchan, M. Bhatt, Robust natural language processing - combining reasoning, cognitive semantics, and construction grammar for spatial language, in: S. Kambhampati (Ed.), Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, IJCAI/AAAI Press, 2016, pp. 2908–2914. URL: http://www.ijcai.org/Abstract/16/413.

[34] M. Spranger, J. Suchan, M. Bhatt, M. Eppe, Grounding dynamic spatial relations for embodied (robot) interaction, in: D. N. Pham, S. Park (Eds.), PRICAI 2014: Trends in Artificial Intelligence - 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings, volume 8862 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 958–971. URL: https://doi.org/10.1007/978-3-319-13560-1_83. doi:10.1007/978-3-319-13560-1_83.

[35] P. R. Cohen, J. Morgan, M. E. Pollack, Intentions in Communication, The MIT Press,

2003. URL: https://doi.org/10.7551/mitpress/3839.001.0001. doi:10.7551/mitpress/3839.001.0001.

[36] L. Dissing, T. Bolander, Implementing theory of mind on a robot using dynamic epistemic logic, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, ijcai.org, 2020, pp. 1615–1621. URL: https://doi.org/10.24963/ijcai.2020/224. doi:10.24963/ijcai.2020/224.

[37] T. Bolander, N. Gierasimczuk, Learning to act: qualitative learning of deterministic action models, J. Log. Comput. 28 (2018) 337–365. URL: https://doi.org/10.1093/logcom/exx036. doi:10.1093/logcom/exx036.

[38] S. Baez Santamaria, T. Baier, T. Kim, L. Krause, J. Kruijt, P. Vossen, Emissor: A platform for capturing multimodal interactions as episodic memories and interpretations with situated scenario-based ontological references, in: Proceedings of the First workshop Beyond Language: Multimodal Semantic Representations in conjunction with IWCS 2022, 2021.

[39] B. J. Grosz, S. Kraus, D. G. Sullivan, S. Das, The influence of social norms and social consciousness on intention reconciliation, Artif. Intell. 142 (2002) 147–177. URL: https://doi.org/10.1016/S0004-3702(02)00274-6. doi:10.1016/S0004-3702(02)00274-6.