Inherently Interpretable Knowledge Representation for a Trustworthy Artificially Intelligent Agent Teaming with Humans in Industrial Environments

Vedran Galetić[1], Alistair Nottle[1]

[1]*Airbus Central R&T, Bristol, United Kingdom*

Abstract

Embodied artificially intelligent agents teaming with humans in industrial environments must be safe and trustworthy, their behaviour predictable, and their rationale explainable. In addressing these extremely wide requirements, we take the knowledge representation angle. Adopting Gärdenfors's Conceptual Space framework, learnt concepts are represented as regions across inherently interpretable quality dimensions, while classification of instances proceeds using a simple derivative model assuming fuzzy category membership. In our use case from the manufacturing domain, the quality dimensions consist of physical properties retrievable from the agent's sensors and utilisation properties from crowdsourced commonsense knowledge. This heterogeneous property decomposition approach allows for flexible concept acquisition and manipulation, particularly useful for industrial settings often characterised by highly specific artefacts and thus data scarcity, which may impact the effectiveness of the state-of-the-art data-hungry — and typically opaque — computer-vision based approaches.

Keywords

Knowledge Representation, Conceptual Spaces, Trustworthy AI, Human-AI Teaming, Human-centred AI, Explainable AI, Embodied Intelligent Agents

1. Introduction and Context

Artificial intelligence (AI) is used across various industry domains, where it achieves state-of-the-art performance in various specific applications, approaching or surpassing human-level performance on cognitive tasks (e.g., [1, 2]).

Sometimes AI is employed in applications where its impact on human safety must be considered, such as embodied AI agents cooperating with humans in performing physical tasks and AI embedded on board an aircraft or a spacecraft. This consideration is particularly relevant for the aerospace domain, where AI may be used across multiple stages of the products' life cycle, most notably manufacturing and operations stages, for example:

- Manufacturing
  - Optimisations by means of scheduling resources for minimising makespan (the time taken to work from start to finish);
  - Automated visual inspection of components and anomaly detection;

AIC 2022, 8th International Workshop on Artificial Intelligence and Cognition
vedran.galetic@airbus.com (V. Galetić); alistair.nottle@airbus.com (A. Nottle)
0000-0003-3861-0202 (V. Galetić); 0000-0001-8767-809X (A. Nottle)

  - Planning task execution for embodied AI agents cooperating with humans on the shop floor;
  - Naturally cooperating with humans using natural language for receiving commands and reporting events;
  - Hybrid modelling of systems combining first principles from domain theory and data-driven optimisation approaches;
- Operations:
  - Cognitive assistance and natural language interaction in the cockpit to ease the pilot's task and increase safety;
  - Automated decision making in fleet management;
  - Space and defence operations involving extremely reliable semantic interpretations of images and videos.

High-performing AI systems, often based on deep neural network models (deep learning), tend to be black boxes in that their internal knowledge, rationale, and decision making may not be readily interpretable to the human users and subjects. Even based on this very limited set of examples of AI usage in the aerospace industry it is clear that the impact of AI systems' decisions may be catastrophic if misdesigned. Thus, a question arises of AI trustworthiness and certifiability. These questions have been a matter of extensive research (e.g., a multi-year consortium programme run by DARPA [3]) and legislation processes conducted by top-level governing bodies in aerospace and other industries, as well as society in general.

Explainable AI (XAI) is a discipline within the AI research field recognised as one of the cornerstones of AI trustworthiness. In fact, the European Commission's High Level Expert Group on Artificial Intelligence identifies explainability (or explicability) as one of the four ethical principles fundamental for trustworthiness of AI, the other three being respect for human autonomy, prevention of harm, and fairness [4]. Furthermore, the Group specifies seven key requirements that are to be addressed throughout an AI product's lifecycle, from which 'transparency', 'accountability', and 'human agency and oversight' are ones clearly related to explainability. These guidelines are upheld by the European Union Aviation Safety Agency (EASA), focussing on certifiability challenges that black-box AI models impose, echoing explainability as one of the three main components of trustworthy AI [5], and fully recognising human-centricity in its AI Roadmap [6].

Endowing an AI system with explainability implies providing interpretable explanations of the system's rationale and outputs [7, 8]. Explanation appropriateness [9] depends on the application context (e.g., time criticality, computing resource constraints) and the role of the human user [10, 11]. Namely, a developer will have different requirements from explainability to help them debug a system from requirements of a pilot, who may need to understand the AI system's recommendation in a very time-critical manner, or a certifier, who may have weeks at their disposal to properly understand the system's decision making rationale and endorse it for use in operation.

Much of the extant work in XAI focuses on post-hoc explainability, i.e., providing explanations of an already existing black-box AI system *a posteriori*. Some of the more popular techniques in this family of techniques include, among others, feature importance analysis

(e.g., SHAP [12], LIME [8]), saliency maps [13], and surrogate modelling by using a simpler, inherently interpretable model, like linear regression or decision trees, to explain the system's behaviour.

On the other hand, taking explainability into account in the design phase of the AI system's life-cycle leads to them being more intrinsically (or inherently) explainable. It is worth noting that we do not consider intrinsic explainability as equivalent to algorithmic or model transparency (e.g., [7]). Instead, in the current work we take an approach towards more intrinsically explainable AI by focusing on the artificial agent's knowledge representation and its interpretability and understandability for the human. We model the agent's knowledge by complementarily combining information obtained by the agent's sensors and openly available general knowledge sources to achieve a high level of knowledge representation flexibility. This will be of particular importance in the settings where pure statistical learning may struggle to capture relevant concepts due to scarcity of available data for rare and specific objects, and in turn to provide a human-understandable account of its rationale and predictions.

2. Knowledge Representation and Inference of (Embodied) AI Agents

An AI agent interacting with human operators or users, particularly if embodied, must be aware of its environment to ensure safe operation. For instance, it should be able to detect and interpret physical objects and categorise events taking place in its surroundings. Interpretability of the embodied AI agent's knowledge content and acquisition processes arguably affords higher trustworthiness of such a system and facilitates its certifiability.

When modelling an artificial agent's knowledge, a classical, long-standing dilemma is encountered regarding the trade-off between the connectionist (neural) and symbolic knowledge representation. The neural modelling excels in learning from statistical regularities inherent in the input data, while these models struggle with concept combinations and interpretability of the resulting knowledge representation. The symbolic view models knowledge using interpretable symbols, amenable for formal logical calculations and syntactic manipulations generating concept combinations, while struggling with mechanistic modelling of concept acquisition processes.

As a naturally intelligent mind indisputably excels at all the above qualities, AI has been in search of the right level of knowledge representation that incorporates and successfully models these qualities in a theoretically sound and computationally amenable manner.

2.1. Conceptual Level of Knowledge Representation

The Conceptual Space theory [14, 15] proposes the conceptual level of knowledge representation, which is situated between the symbolic and neural representation levels. It adopts basic tenets of cognitive semantics, thus acknowledging that the meaning itself is realised within the agent's reasoning system rather than objective and external to the agent. It takes into account empirically founded phenomena such as prototype-based concept organisation [16, 17, 18] and schematicity [19].

A conceptual space is a geometric space spanned by quality dimensions. Quality dimensions group into integral groups constituting quality domains. Examples of quality domains are *colour*, *size*, and *taste*, while examples of integral quality dimensions are *hue*, *saturation*, and *brightness* (of *colour*). Properties (e.g., *green*) are convex regions of quality domains, concepts (e.g., *apple*) are intersections — again, convex — of relevant properties, while individual objects are points in the space, characterised by specific values of pertinent quality dimensions. When properties to describe a concept are languageable, it is possible to remark that such a conceptual space represents an inherently interpretable conceptual knowledge representation of a reasoning agent.

As this representation is geometrically organised, it is possible to compare objects and quantify their similarities employing well-established distance metrics, such as Euclidean or Manhattan. Comparison of instances in the space of quality dimensions overcomes a typical pitfall in pure distributional semantic approaches, where it is sometimes a challenge to discern between similarity (e.g., *aeroplane–rocket*) and relatedness (e.g., *aeroplane–pilot*). Concept combination is modelled by overlapping space regions, while space transformation operations are defined for accounting for metaphor and metonymy. It is worth noting that, taking into account the cognitive motivation of the theoretical framework, quality dimensions are seen as essentially cognitive constructs that model the conceptual structure consistently with the cognitive semantics tenets. Euclidean space may sometimes not be the most appropriate model; instead, the polar coordinate system is sometimes more suitable (e.g., for *colour*; see [14]). Consequently, when operationalising the framework special attention should be paid to choosing the correct set — and structure [20] — of quality dimensions, bearing in mind some may not be languageable, especially in case of abstract concepts (e.g., *love*, *idea*, *strength*).

In the context of AI trustworthiness, the Conceptual Space framework adds to interpretability in multiple ways. On the one hand, a conceptual space can be seen as an interpretable knowledge representation level as the space itself is typically spanned by languageable (or at least interpretable) quality dimensions. This should be of high usability for a developer of such a system for debugging it and identifying its vulnerabilities, or an examiner (e.g., certifier) who will want to understand the AI agent's contained knowledge as input to its endorsement for operation in an environment involving humans. On the other hand, it can be viewed as an interpretable 'checkpoint' in neuro-symbolic mapping. Namely, sensory information is abstracted to the geometrically organised space via pertaining quality dimensions, while arbitrary symbols are grounded onto concepts represented as convex regions of the space.

There are already a few implementations of Conceptual Spaces (e.g., [21]) as well as quite a few applications in various domains [22]. While some of them proved promising for particular use cases, we found a previously conceived simple model for typicality quantification serves the current use case adequately.

3. Classification and Typicality Model

We draw inspiration from Gärdenfors's original framework and propose to use a simple model aimed at formalising interpretable classification of artefacts in the manufacturing domain. We employ the property decomposition approach in representing concepts and assume the graded

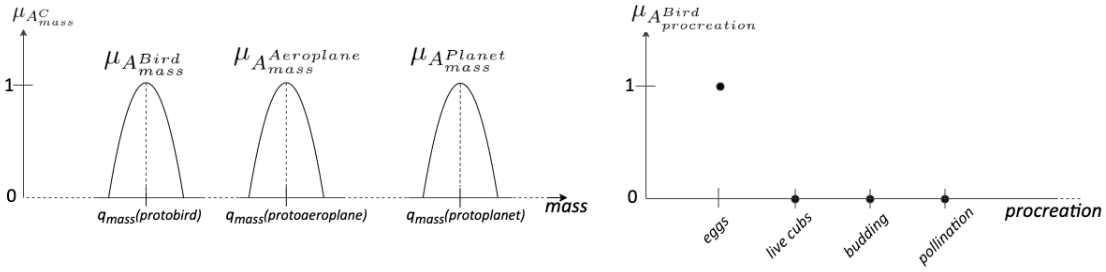

Figure 1: Left: Illustration of the membership functions μ for a continuous quality dimension $mass$ across three example concepts. Prototypical instances are assumed to bear prototypical value of the property. Right: Illustration of a nominal quality dimension $procreation$ for the $Bird$ concept.

membership of their instances, i.e., we do not adopt the Aristotelian view on categorisation governed by necessary and sufficient conditions; instead, we represent categories as fuzzy sets [23]. Therefore, instances are members of categories to various degrees expressible by a real-valued membership function, while also taking into account different property weights.

Concretely, the model, called *μw-model* [24], is based on the prototype theory [16]. It uses two parameters to classify an instance (e.g., a concrete object) and quantify its typicality: the membership function (μ) and the property weight (w). The membership function μ is defined for distribution of values for each quality dimension of a concept, while the property weight w is defined per quality domain for a specific concept.

Fig. 1 illustrates the membership function μ for a continuous and a nominal quality dimension. Its distributions clearly indicate how typical an instance would be with respect to values of each of these properties. For example, a dolphin instance is a typical *fish* with respect to *shape*, but an untypical one with respect to *procreation* (Fig. 1 right). In fact, this exemplifies the impact of lexical definition constraints and theories that may supersede similarity-based classification (see, e.g., [25]). These constraints are learnt explicitly with the aim of consensual definitions amongst the community for efficient communication [26] and are beyond the scope of the current work focussing on a limited domain.

The weights w allow us to quantify the importance of different quality domains for the typicality of instances across concepts. For example, *colour* is an important diagnostic property for many natural kinds [17], especially those that do not have a high variability thereof across the specimens; conversely, *colour* tends to be less conceptually central (e.g., [27]) for artefacts since many can be painted arbitrarily.

Both of these parameters are learnt from the agent's supervised experience in the environment, which entails obtaining labels of instances from the domain expert. Concretely, the distributions over property values utilised by the μ parameter are learnt in a frequentist manner across labelled instances characterised by quality dimension values extracted from own sensors or retrieved from knowledge bases (see § 4). The property weights w pondering these properties in classification are learnt in a theory-based manner, using causal status hypotheses [28, 29]. For instance, relying on the notion of conceptual centrality [27], it has been empirically proven that the property weight is inversely proportional to its mutability and variability across concept instances [30]. For example, *colour* is more conceptually central (thus bears higher weight

w) for *orange* than for *apple* because the *orange* concept has lower mutability and variability of *colour* across its instances than the *apple* one. Formal systematic quantification of the *μw-model*'s parameters are beyond the scope of this paper; however, for theoretical development and empirical findings please consult the above references.

The *μw-model* allows for quantification of an instance's typicality with respect to a concept. Specifically, an instance is represented as a vector whose dimensions are the quality dimensions spanning the conceptual space, while each coefficient is the membership function value pondered by the root of the normalised weight, namely:

$$\overrightarrow{r_C}(c) = \sum_i \sqrt{\frac{w_i(C)}{\sum_j w_j(C)}} \mu_{A_i^C}(q) \cdot \overrightarrow{e_i} \tag{1}$$

where i denotes a quality dimension (e.g., height), $\overrightarrow{e_i}$ are basis vectors spanning the space, j is a quality domain (e.g., *taste*), q is an instance, q_i is the value of the quality dimension i for that instance (e.g., *51 cm*), $w_i(C)$ is the weight of the quality dimension i for the concept C, and $\mu_{A_i^C}(q)$ is the typicality measure of the instance q for the concept C with respect to the quality dimension i (e.g., the representativeness of '*this tiger-striped ball*' for the concept *tiger* with respect to the quality dimension *texture*; which would be high for this dimension, but extremely low for virtually any other dimension).

The typicality of an instance in the frame of the context is represented as the second norm of the vector from Eq. 1, namely:

$$R_C(c) = \|\overrightarrow{r_C}(c)\| = \sqrt{\sum_i \frac{w_i(C)}{\sum_j w_j(C)} \mu^2_{A_i^C}(q)} \tag{2}$$

The concept for which its representativeness (Eq. 2) is the highest is calculated in the straightforward way:

$$Cat(c) = \max_C R_C(c) \tag{3}$$

Such non-binary classification output allows for exploration into the predictor's rationale and intrinsically interpretable factors that influenced the prediction.

4. Crowdsourced Knowledge Component

Humans possess a vast amount of reusable knowledge components — based on the commonsense knowledge — consisting of empirically acquired high-level hypotheses imposing constraints on the interpretation of situations and tasks at hand [31, 20]. Understandably, such knowledge is elusive for AI agents and its acquisition and formulation is anything but trivial. Without it, an agent is left with imperfect priors to rely on, stemming from its modest experience with the very constrained domain environment it is typically embedded in.

Fortunately, there are openly available knowledge bases comprising general and connotative knowledge in a structured form, thus computationally amenable for the AI agent. These knowledge bases are typically ontologies [32], hierarchies [33], or other types of knowledge

graphs. They are generally obtained through manual or semi-automatic specification of concepts and their interrelationships, sometimes with the aid of crowdsourcing.

ConceptNet [34] is an openly available resource comprising general knowledge obtained from extant manually crafted knowledge bases (e.g., Wiktionary, OpenCyc [32]) and dedicated large-scale crowdsourcing campaigns. The resulting multilingual knowledge graph contains 8 million nodes (concepts — words or syntagms) and 21 million edges (relations between concepts). What makes it particularly interesting is that there are only 36 different relations, making the resource semantically parsimonious as well as reasonably tidy and manageable. Authors have validated the soundness of ConceptNet's inherent semantics by combining then-state-of-the-art word embeddings with ConceptNet using a modified version of the so-called retrofitting technique [35], demonstrating that distributional semantics endowed with input from general knowledge graphs yield higher performance on a standard linguistic task of word relatedness quantification.

Given the context of human-machine interaction in an operational environment such as manufacturing, incorporating commonsense knowledge (concretely, about artefact utilisation) is a crucial input for modelling the knowledge module of an AI agent teaming with the human operator. This is particularly important in use cases where artefacts are quite specific and training the agent to classify objects based solely on computer vision-based machine learning techniques may be difficult due to insufficient training data, and mistaking an object's purpose due to incorrect similarity-based object classification may prove detrimental.

ConceptNet's knowledge contained among relations such as *UsedFor*, *MadeOf*, or *PartOf* addresses this very problem successfully. These properties will have significant weights in the context of representativeness measurement using the *μw-model* (see § 3).

5. Manufacturing Domain Use Case

Whilst robotics are not unfamiliar within a factory setting, they are often relied upon for repetitive tasks, such as drilling, or assembly. Their strength and precision make them ideally suited to these highly constrained applications, and are showing key improvements in productivity [36]. In the mid-term future it is expected that humans shall be accompanied by embodied artificially intelligent agents, which will team with humans in performing various tasks, in order to optimise execution times and reduce non-productive time, for example, to make hand-held tools available where needed and when needed, to assist the human in repetitive tasks, to alert the human of possible deviations and risks in execution, *etc.* More recent developments in robotics, and human-robot interaction, allow additional utility through the use of collaborative robots, or 'cobots'. These 'cobots' will be able to assist in faster, safer, and higher quality completion of industrial activities.

To extend the environment awareness of cobots, we need to ensure recognition of new objects that both offer powerful performance, as well as clear, understandable explanations for the outcomes of machine classifications. We model the artificial agent's knowledge interpretably using the Conceptual Space framework [14] to ensure performant recognition of previously unseen objects, while at the same time helping to build trust and confidence in the system (from a system developer, certifier, and operator perspective) and reduce cognitive loadings of

	Colour	Texture	Shape	Composition	UsedFor
$\mu_{A_i^{Drill}}(new\_object)$					
$w_i(Drill)$	LOW	LOW	MEDIUM	MEDIUM	HIGH

Figure 2: The artificial agent wants to classify a new object (upper right image) and extracts its physical and utilisation properties. The table (below) illustrates the μ and w values for the 'Drill' concept. Although the new object is physically similar to the learnt 'Drill' prototype (upper left image shows a typical instance of 'Drill'), it will not be classified as one due to incompatibility of the utilisation property bearing the highest weight.

explanation consumers. For instance, a system which can classify something as a drill because it has a similar size, colour, and shape as previously seen drills, and, crucially, utilisation properties indicating the unseen item is used for drilling, allows for an interpretable classification from the system — i.e., '*I believe this is a drill as it looks similar to other drills I've seen in the past, and it is used for drilling*'. Conversely, an object that bears surface similarities to previously seen drill instances yet is perceived to be used for riveting should instead be classified as a riveter, reflected by expectation that utilisation properties of industrial artefacts are more important for their classification than their surface properties (Fig. 2).

This has the additional benefit of helping tackle the data scarcity problem — particularly in environments which are non-typical — where data-hungry computer vision algorithms are starved of their normal surfeit of images.

5.1. From Object Detection to Property Decomposition

Within the context of our research, the current industrial setup does not readily allow for deployment of cobots on the factory floor, so we have instead relied on a simulated environment to provide training data and to gain feedback and validation of the proposed methodologies. The Webots Open Source Robotics Simulator [37] was used to create, in the first instance, a simple 'playground' environment with a controllable e-puck robot [38] equipped with a simple sensor package, including standard vision sensor (i.e., a camera) as well as a time-of-flight sensor.

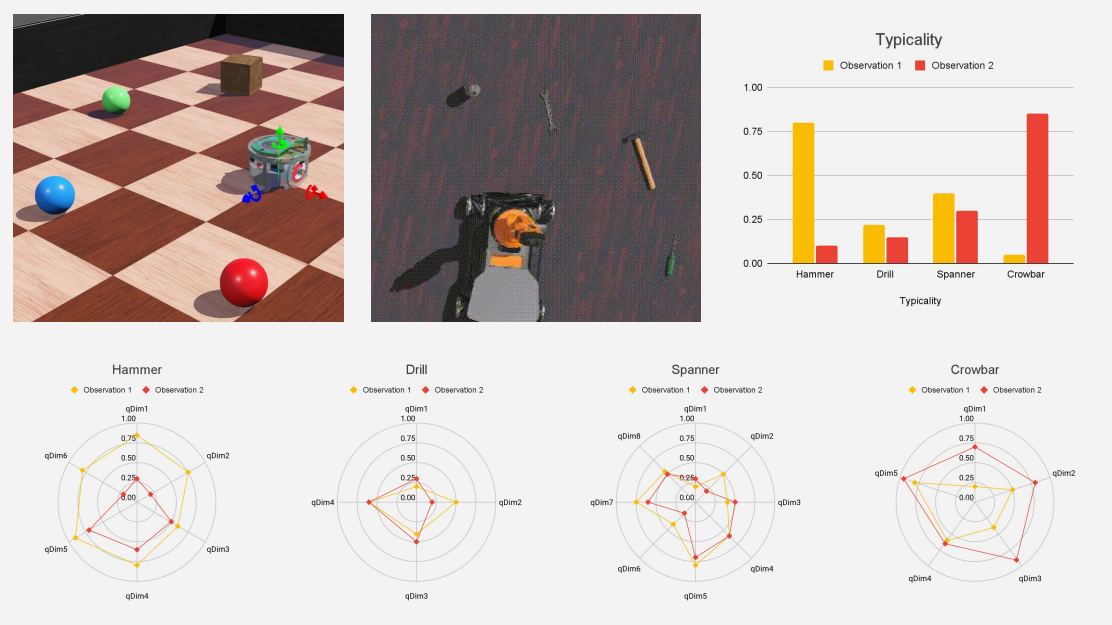

Figure 3: Simulated industrial use case environment. We start with the simple case with simulated idealised objects (upper left), such as 'Red Sphere', 'Wooden Cube', *etc.* The robot encounters these objects treated as prototypes and extracts ground-truth property values available from the simulator ('colour', 'shape', 'composition'), used as the basis for learning the conceptual space by Voronoi tessellation. Having validated the data pipeline and knowledge representation modelling, we moved towards the more ecologically valid manufacturing simulated environment (upper middle), where quality dimensions include physical properties like 'size', 'texture', 'composition', and utilisation properties. Generally, every instance's quality dimension values are used as input for membership quantification per various concepts of interest. The spider charts (lower) show example membership values per quality dimensions across four artefact concepts for two example observations, one in orange, the other in red. Together with the quality domain weights (currently arbitrated, in the future quantified via empirical hypotheses, see § 3 and § 6), these membership values make up the vector representation of the instance, as per Eq. 1. The bar chart (upper right) represents the typicality of the two example instances across the four observed concepts, calculated via Eq. 2.

Throughout the simulation environment are a number of different objects. In the first iteration of the environment (Fig. 3), these were simplistic idealised example objects, such as a 'Green Ball' or 'Red Cube', to allow for 'easier' recognition. As the environment develops, additional objects are introduced, which are either 'off the shelf' or custom created, to represent real-world, ecologically relevant, objects such as a hammer or screwdriver.

Within the simulation environment we attach additional custom properties to the objects (e.g. stripy texture). That way ground truth values can be gained from the simulation environment, such as colour, texture, utilisation properties, *etc.* This, along with the provided labels of the encountered objects, provides a baseline from which to learn the conceptual representation. Concretely, a conceptual space is learnt by Voronoi tessellation [39] around prototypes[1]. In the

[1]It is worth noting that, apart from prototype-centric categorisation, literature also suggests exemplar-based concept

simple introductory use case the idealised objects are considered to already be prototypes in order to expedite the proof-of-concept space construction, while in the more complex cases prototypes are calculated as centroids of the acquired labelled instances at training time.

As the robot is controlled throughout the world, the field-of-view of the camera is calculated, and any objects which come into sight then trigger the robot controller to submit the current image to our API, alongside the ground truth properties relating to the object.

During real-world implementation, we would anticipate that the capturing of images and detection of objects would be carried out using state-of-the-art object detection algorithms. However, as these are not the focus of our research, we utilise ground truth data and imagery directly from the simulation environment, while the property detectors are a matter for other active AI research . This has the added benefit of allowing the user to select the most appropriate algorithm for their particular use case.

Once submitted to the API, a number of property detectors are applied to extract properties of the observed object. These take two broad categories of detectors: physical property detectors; and utilisation property detectors. Basic physical properties can be inferred from the sensors on the robotic platform. Examples of properties we have experimented with include:

- Texture – using Concept Activation Vectors [41] to determine distinctive textural properties of an object, e.g. stripey, smooth, etc.;
- Colour – using simple computer vision approaches to determine the dominant colour of the object. Colour is represented within the HSB colour space;
- Shape – building on work by Lucas Bechberger to define the shape of the object with just a few properties [42];
- Size – determined by a depth-aware camera.

Utilisation properties, evocative of *affordances* [43], are properties which require additional transformation to derive. Sources for these could include:

- Task recognition through computer vision techniques to determine which tasks (if any) are being carried out in the observed video stream, e.g. 'hammering' or 'drilling';
- Crowdsourced knowledge base, providing information on uses of particular entities within its knowledge graph.

In the current stage of our research we use the latter for extracting the utilisation properties. Concretely, we use ConceptNet's (see § 4) API[2] to acquire crowdsourced values for the *UsedFor* property of various manufacturing artefacts (e.g., *drill*). Each response (e.g., *'drilling holes in things'*) is accompanied by the weight, reflecting the reliability of the source. Due to the nature of free-form submissions, messiness of the corresponding response cannot be avoided (e.g., there is a separate response entry for '*drilling holes in things*' and '*drill a hole in something*'), which means we cannot copy the weight values to our μw-model context. Instead, we group all responses with similar semantics, such as the examples above, and run the softmax function

organisation [40], which may be a matter of future work.

[2] http://api.conceptnet.io (accessed on 23 March 2022)

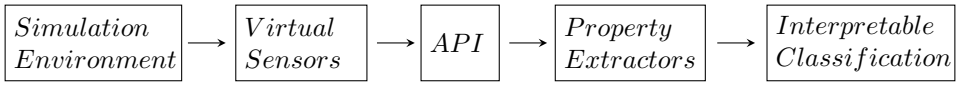

Figure 4: High-level data flow from simulated environment to interpretable classification system.

across all weights of semantically different items to acquire the quantity that we use to ground the membership function[3] for this utilisation property. Concretely, we get

$$\mu_{A^{Drill}_{UsedFor}}(this\_object\_being\_used\_for\_drilling\_holes) = 0.9999997$$

which is a reasonable quantity. This and other quality dimensions' membership values, along with dimension weights, make up the representativeness vector (Eq. 1), from which we measure the instance's typicality across concepts (Eq. 2) and determine in which one its representativeness is the highest (Eq. 3, visualised in Fig. 3).

Through this approach we aim to move from an object detection approach to a property detection based approach, and provide a sample pipeline (Fig. 4) for how this can be implemented within a simulated environment. The majority of this pipeline, from the API onward, will then allow for the implementation of state-of-the-art property detectors. Similarly, the simulation environment could be substituted with real-world cobots with real-world sensors maintaining the same interfaces throughout.

5.2. Qualitative Validation

As an industrial implementation, we have had to seek validation from future end users across a number of different stakeholders, showing that the system has a grounding in reality and could form the basis of a deployed industrial system. Validation must be sought at a strategic level — that our proposal fits in with the strategic direction being taken by the company. In our context, this has involved engaging directly with technical roadmap owners, who have strategic oversight of AI and Robotics. Feedback from these stakeholders confirms that our approach is compatible with a vision of the future industrial set up.

Consideration has also been given to how classification results are returned to users in a meaningful way. Fortunately, the inherently explainable nature of the classification process can make simple explanations relatively easy, e.g. *'This is a drill because it is the right size and shape and is used for drilling'*. Visualisations can also be of assistance. One approach taken is the use of a spider (or radar) chart (see Fig. 3). Each of the axes (from 0 to 1) represents one the dimensions of a concept, with 0 being 'not at all typical' to 1 being 'prototypical'. Therefore, it is possible to map different observations on this space and the larger the area of overlap between the observation and the prototype the higher the typicality of the instance with respect to that class. A user can have a quick insight into how 'well matched' the observation is and gain confidence in the system.

[3]Somewhat counterintuitively, what is called 'weight' in the ConceptNet system is semantically closer to the membership function μ of the *μw-model* than the weight parameter w. See § 3 for details.

6. Conclusion and Future Work

We focus on (embodied) artificially intelligent agents teaming with humans in industrial environments. In this context, we propose a flexible knowledge representation building block aimed at increasing the agent's environment awareness, its operational predictability, and explainability of its rationale, ultimately leading to increased trustworthiness of the AI system, thus paving the way to its certifiability.

We pursue a heterogeneous knowledge modelling approach, relying on physical properties acquirable by equipped sensors and utilisation properties obtained from openly available crowdsourced commonsense knowledge bases. These properties are represented on a common representation model drawing inspiration from Gärdenfors's Conceptual Space framework. The model is defined by two parameters: one is the membership function in the context of fuzzy set theory (quantified frequentistically across different properties); the other represents the weight of the property (quantifiable based on empirically confirmed hypotheses in the area of cognitive semantics, which is a matter of upcoming work).

Moving from computer vision-based object recognition to interpretable classification engendered by the property decomposition approach of entity representation is particularly relevant and useful in applications characterised by data scarcity, i.e., when not many images of highly specific objects exist (as is the case in our industry) to train a classifier based on (convolutional) neural networks. Apart from flexibility, inherent interpretability of components constituting the knowledge representation formalism allows for model inspection by a developer, rigorous examination by a certifier, and output understandability for an operator using or cooperating with the AI.

Authors fully recognise that the described work represents a limited use case, which opens many avenues for future research in the context of knowledge representation and inference modelling in (embodied) AI agents teaming with humans, and we need to call on the latest research in the field (e.g. [44]).

Humans as intelligent agents are remarkably successful in learning concepts from very scarce data [45]. The mechanics and phenomenology of it is of high interest to cognitive science, cognitive neuroscience, computational neuroscience, and artificial intelligence [45, 46]. The challenge is to model effortless and reliable processing leading to acquisition of core concepts and accompanying intuitive theories (e.g., in physics and psychology) as well as generic causal structures [47]. Unsupervised learning approaches, like autoencoders, are very relevant when it comes to knowledge representation and concept learning. A successful model contains a compressed representation of environment phenomena still retaining the right information necessary for a faithful reconstruction. Some promising generative models are based on Bayesian reasoning in the context of hierarchies and structures of hypotheses and associated inductive constraints [20]. An artificial agent's successful acquisition and manipulation of core concepts arguably makes its behaviour more predictable, its rationales more interpretable, and it itself more trustworthy.

Generic causal structures of hyper-hypotheses governing the acquisition and manipulation of the core concepts give rise to interdependencies among quality domains, which has been omitted from this paper so far, instead representing quality dimensions via orthogonal basis vectors. Property correlations are an important part of concepts' structures (e.g., [48]) as it

has been empirically demonstrated that people effortlessly acquire systematic correlations [49, 50, 51, 52]. For example, the colour of fruit is an indication of its taste and ripeness. These covariations are particularly immanent to natural kinds and pertaining theoretical developments deal with causality and hint at the philosophical notion of psychological essentialism, stating that observable features are only a guidance to the true nature of objects [53, 54]. For artefacts, which can arguably take arbitrary property values, it makes sense to focus attention on affordances stemming from an object's physical properties, which is of particular relevance for the industrial domain (e.g., a large hand-held object with a steel flat tip is likely used to hit nails).

Apart from (physical) objects, newer work in the Conceptual Space theory looks at modelling events via force and result vectors [55, 56]. It is a promising research stream and particularly relevant for the current industrial domain with human-AI teaming. An interesting challenge to tackle will be interpretability evaluation of quality dimensions representing force patterns that the event-based extension of the Conceptual Space framework suggests.

An interpretable representation of objects and events would make a good basis for a declarative knowledge module of a cognitive architecture [57, 58] used for cognitive modelling typical for scenarios involving cognitive assistance (e.g., [59]), particularly task modelling, behaviour deviation detection, and cognitive load quantification. While using subsymbolic formalisms for declarative knowledge representation is not novel (e.g., [25, 60]), their utilisation and validation are yet to be demonstrated in industrial environments and highly critical applications involving AI.

Finally, it is important to reiterate that explainability, albeit undeniably important in human-centric applications, is but one pillar of AI trustworthiness and certifiability. Responsible and ethical design of AI is a *sine qua non* for such use cases. Other notable research areas compatible with explainability are robustness and learning assurance, and fairness and non-discrimination [4, 5, 6]). Clearly, the path towards AI trustworthiness is complex and multi-faceted, and will be difficult to address without interdisciplinary research, ideally conducted jointly by industry and academia.

References

[1] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[2] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, A. A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, Nature medicine 24 (2018) 1716–1720.

[3] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI magazine 40 (2019) 44–58.

[4] European Commission High-level Expert Group on Artificial Intelligence, Ethics guidelines for Trustworthy AI, European Commission, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[5] EASA, EASA Concept Paper: First usable guidance for Level 1 machine learning applications, EASA, 2021. URL: https://www.easa.europa.eu/sites/default/files/dfu/easa_

concept_paper_first_usable_guidance_for_level_1_machine_learning_applications_-_proposed_issue_01_1.pdf.
[6] EASA, EASA Artificial Intelligence Roadmap 1.0 A human-centric approach to AI in aviation, EASA, 2020. URL: https://www.easa.europa.eu/document-library/general-publications/easa-artificial-intelligence-roadmap-10.
[7] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57.
[8] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[9] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[10] R. Tomsett, D. Braines, D. Harborne, A. Preece, S. Chakraborty, Interpretable to whom? A role-based model for analyzing interpretable machine learning systems, arXiv preprint arXiv:1806.07552 (2018).
[11] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle, A. Preece, A systematic method to understand requirements for explainable ai (xai) systems, in: Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China, volume 11, 2019.
[12] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[13] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).
[14] P. Gardenfors, Conceptual spaces: The geometry of thought, MIT press, 2004.
[15] P. Gardenfors, The geometry of meaning: Semantics based on conceptual spaces, MIT press, 2014.
[16] E. Rosch, B. B. Lloyd (Eds.), Cognition and Categorization, L. Erlbaum Associates, 1978, pp. 27–48.
[17] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cognitive psychology 8 (1976) 382–439.
[18] E. Rosch, C. B. Mervis, Family resemblances: Studies in the internal structure of categories, Cognitive psychology 7 (1975) 573–605.
[19] G. Lakoff, Women, fire, and dangerous things: What categories reveal about the mind, University of Chicago press, 2008.
[20] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction, science 331 (2011) 1279–1285.
[21] L. Bechberger, K.-U. Kühnberger, A thorough formalization of conceptual spaces, in: Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz), Springer, 2017, pp. 58–71.
[22] F. Zenker, P. Gärdenfors, Applications of conceptual spaces, Cited on 25 (2015).
[23] L. A. Zadeh, Fuzzy sets, in: Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh, World Scientific, 1996, pp. 394–432.
[24] V. Galetić, An aggressive robin in the backyard: Formal quantification of prototypicality level within the frame of the prototype semantic theory of cognitive linguistics, Suvremena

lingvistika (Contemporary Linguistics) 37 (2011).
[25] A. Lieto, D. P. Radicioni, V. Rho, Dual peccs: a cognitive system for conceptual representation and categorization, Journal of Experimental & Theoretical Artificial Intelligence 29 (2017) 433–452.
[26] V. Galetić, Z. Jelaska, Typification and formal quantification of prototypical in language acquisition and learning, Lahor: časopis za hrvatski kao materinski, drugi i strani jezik 1 (2011) 39–64.
[27] S. A. Sloman, B. C. Love, W.-K. Ahn, Feature centrality and conceptual coherence, Cognitive Science 22 (1998) 189–228.
[28] W.-k. Ahn, N. S. Kim, M. E. Lassaline, M. J. Dennis, Causal status as a determinant of feature centrality, Cognitive Psychology 41 (2000) 361–416.
[29] W.-k. Ahn, S. A. Gelman, J. A. Amsterlaw, J. Hohenstein, C. W. Kalish, Causal status effect in children's categorization, Cognition 76 (2000) B35–B43.
[30] V. Galetić, Formalisation and quantification of a cognitively motivated concecptual space model based on the prototype theory, Ph.D. thesis, University of Zagreb, 2016.
[31] J. B. Tenenbaum, T. L. Griffiths, C. Kemp, Theory-based bayesian models of inductive learning and reasoning, Trends in cognitive sciences 10 (2006) 309–318.
[32] D. B. Lenat, R. V. Guha, Building large knowledge-based systems; representation and inference in the Cyc project, Addison-Wesley Longman Publishing Co., Inc., 1989.
[33] G. A. Miller, Wordnet: a lexical database for english, Communications of the ACM 38 (1995) 39–41.
[34] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Thirty-first AAAI conference on artificial intelligence, 2017.
[35] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, N. A. Smith, Retrofitting word vectors to semantic lexicons, arXiv preprint arXiv:1411.4166 (2014).
[36] M. T. Ballestar, Á. Díaz-Chao, J. Sainz, J. Torrent-Sellens, Impact of robotics on manufacturing: A longitudinal machine learning perspective, Technological Forecasting and Social Change 162 (2021) 120348.
[37] Cyberbotics, Webots: robot simulator, https://cyberbotics.com/, 2022. Accessed: 2022-02-18.
[38] EPFL, e-puck education robot, http://www.e-puck.org/, 2018. Accessed: 2022-02-18.
[39] P. Gärdenfors, M.-A. Williams, Reasoning about categories in conceptual spaces, in: IJCAI, Citeseer, 2001, pp. 385–392.
[40] B. C. Malt, An on-line investigation of prototype and exemplar strategies in classification., Journal of Experimental Psychology: Learning, Memory, and Cognition 15 (1989) 539.
[41] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.
[42] L. Bechberger, M. Scheibel, Representing complex shapes with conceptual spaces, in: Second International Workshop'Concepts in Action: Representation, Learning, and Application'(CARLA 2020), 2020.
[43] J. J. Gibson, The theory of affordances, Hilldale, USA 1 (1977) 67–82.
[44] A. Lieto, Cognitive Design for Artificial Minds, Taylor & Francis, 2021. URL: https://books.google.co.uk/books?id=Xm4ZEAAAQBAJ.

[45] D. Hassabis, D. Kumaran, C. Summerfield, M. Botvinick, Neuroscience-inspired artificial intelligence, Neuron 95 (2017) 245–258.
[46] N. Kriegeskorte, P. K. Douglas, Cognitive computational neuroscience, Nature neuroscience 21 (2018) 1148–1160.
[47] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building machines that learn and think like people, Behavioral and brain sciences 40 (2017).
[48] M. Raubal, Formalizing conceptual spaces, in: Formal ontology in information systems, proceedings of the third international conference (FOIS 2004), volume 114, 2004, pp. 153–164.
[49] D. Billman, J. Knutson, Unsupervised concept learning and value systematicitiy: A complex whole aids learning the parts., Journal of Experimental Psychology: Learning, Memory, and Cognition 22 (1996) 458.
[50] S. S. Jones, L. B. Smith, How children know the relevant properties for generalizing object names, Developmental Science 5 (2002) 219–232.
[51] H. Kloos, V. M. Sloutsky, What's behind different kinds of kinds: effects of statistical density on learning and representation of categories., Journal of Experimental Psychology: General 137 (2008) 52.
[52] J. L. McClelland, T. T. Rogers, The parallel distributed processing approach to semantic cognition, Nature reviews neuroscience 4 (2003) 310–322.
[53] H. Kornblith, Inductive inference and its natural ground: An essay in naturalistic epistemology, Mit Press, 1995.
[54] V. Galetić, Towards the cognitive plausibility of conceptual space models, Suvremena lingvistika (Contemporary Linguistics) 41 (2015) 71–85.
[55] A.-L. Mealier, G. Pointeau, P. Gärdenfors, P. F. Dominey, Construals of meaning: The role of attention in robotic language production, Interaction Studies 17 (2016) 41–69.
[56] P. Gärdenfors, An epigenetic approach to semantic categories, IEEE Transactions on Cognitive and Developmental Systems 12 (2018) 139–147.
[57] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, Y. Qin, An integrated theory of the mind., Psychological review 111 (2004) 1036.
[58] J. E. Laird, C. Lebiere, P. S. Rosenbloom, A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics, Ai Magazine 38 (2017) 13–26.
[59] O. W. Klaproth, M. Halbrügge, L. R. Krol, C. Vernaleken, T. O. Zander, N. Russwinkel, A neuroadaptive cognitive model for dealing with uncertainty in tracing pilots' cognitive state, Topics in cognitive science 12 (2020) 1012–1029.
[60] A. Lieto, A. Chella, M. Frixione, Conceptual spaces for cognitive architectures: A lingua franca for different levels of representation, Biologically inspired cognitive architectures 19 (2017) 1–9.