**PhD Thesis**

**Towards Human-like Robot Perception: Physical Reasoning based on Embodied Probabilistic Simulations for Advanced Robot Manipulation**

*By*

*Franklin Kenghagho Kenfack*

**Keywords:** Physical Reasoning, Perception, Autonomous Robot, Mental Simulation, Commonsense.

1. **Research Motivation**

Developing cognition-enabled robot agents that can competently perform complex human-like everyday manipulation tasks such as cooking in a realistic kitchen has been a worthful challenge not only from a scientific viewpoint but from an engineering viewpoint too. This challenge is particularly being addressed by the collaborative research center EASE CRC[[1]](#footnote-0) in Germany for the EASE project (Everyday Activity Science and Engineering). Regarding this challenge, perception has been a major bottleneckfor autonomous mobile robotics. That is, the perception system should provide appropriate information about the world state for effective and successful manipulation. However, actual perception systems have been struggling against the complexity of such human-centered and dynamic environments, as depicted in figure 1 below, at least in three points. The word *perception* is primarily used to refer to *vision* since it is the main channel for monitoring the external world.

As first issue, **(a)** **sensor information are very limited**. One distinguishes between extrinsic limitations such as clutter, occlusion, noise (e.g., overlay, signal interference and attenuation, embodiement-constrai- ned noise) and intrinsic sensor limitations such as missing depth for smooth (i.e., specular reflection of depth camera rays) (see Figure 1.6) and glass objects (i.e., ray transmission rather than reflection) (see Figure 1.7) or the inability to measure certain physical quantities such as the speed of flow of a liquid when pouring. Attempts to solely rely on these sensor information leads to a situation where compression while learning is no more efficient due to higher entropy in the training dataset.

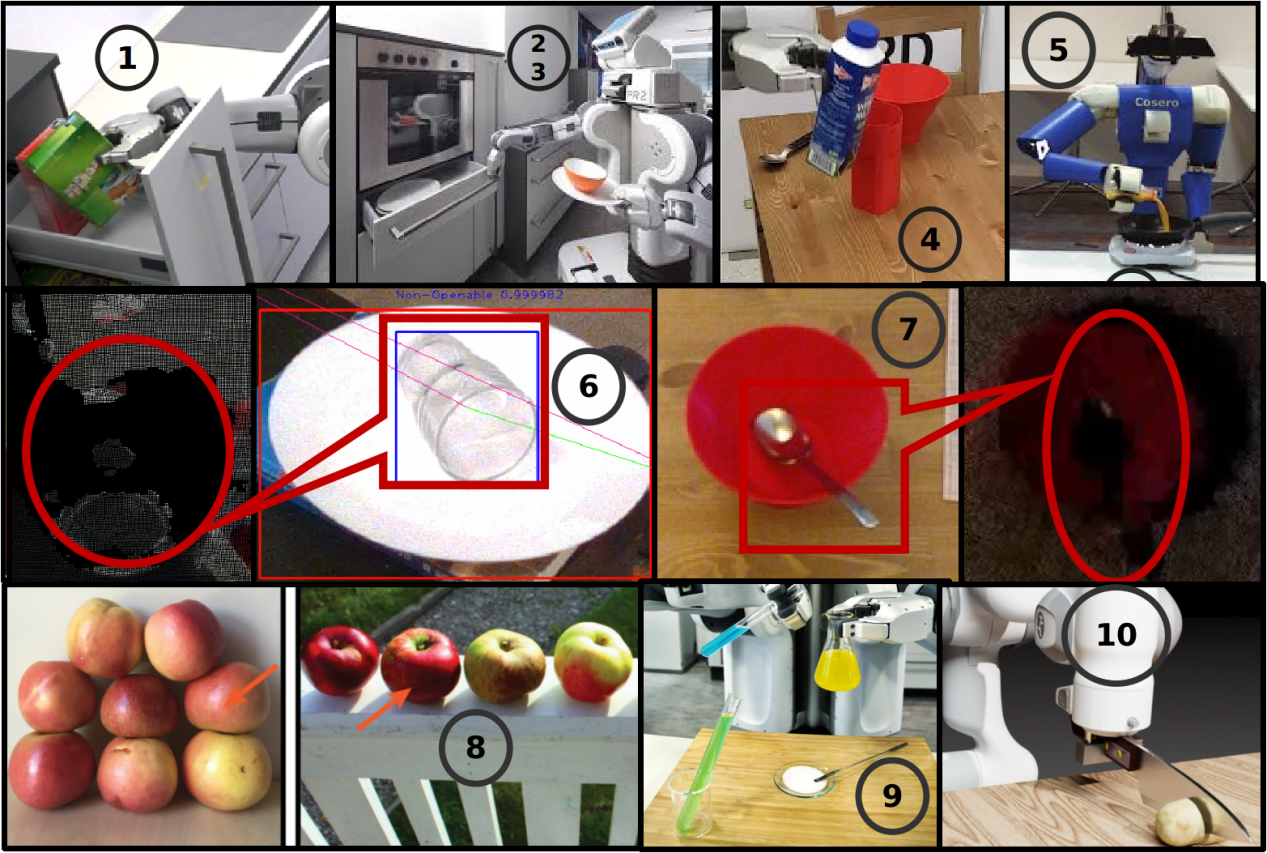


Fig. 1: Illustration of complex robot manipulation situations (1-6)

An attempt to overcome this limitation of sensor information has been the simplification of the environement (e.g., no transparent, fluid, tiny nor overlapping entities, only plane support surface, single-agent assumption) and then the integration of background knowledge about such simplified environment also referred to as priors or common sense knowledge of the world. However, this is not only deceiving w.r.t. (a) but also enforces the staticness of the scene, which is unrealistic. Contrarily, human-centered environments are very dynamic, where entity states are permanently subject to changes and sometimes independently of the agent’s intentions. In this regard, imagine the scenario (2,3) on Figure 1 above, the robot holding a plate containing a bowl and trying to open the drawer, despite the fact that the robot camera is focused on the drawer, the robot should still be aware of the state of the bowl. Another scenario is the case of a robot trying to pour some milk from a bottle into a mug (6). Notice that success depends on the robot understanding of the milk fluid dynamics and how to control it by manipulating the bottle in order to ensure for instance that the milk will neither fall out of the mug, the mug will not spill, nor the mug will be overfilled. There is actually no sensor that would directly support the robot in preventing either one of the above undesired state of the mug or milk. Finally, let consider the scenario (4) where the robot is trying to place a bottle of milk on the table. Though this is a trivial task for humans, note that this is challenging for robots, where the bottle quite often falls instead of standing on the table. Notice that many physical factors come into play when deciding whether or not the bottle will stand after the robot gripper releases it. These are for instance the relief of the table surface, the weight, pose, content (e.g., liquid, powder) and shape of the bottle. These scenarios suggest, as second issue, that perception systems should go beyond addressing classical recognition and localization tasks from actual sensor data, but also reason about the causes and consequences of the agent-object and object-object interactions in order to anticipate (undesired) states of the environments, suggesting therefore a **(b) lack of prospection** in actual perception systems. This essence of prospection in perception is illustrated in Figure 1.8 and 1.1 that shows that recognizing and localizing apples in the scene are one thing, but informing the manipulation system about how to grasp them or which one to grasp is another thing. While the other apples will collapse if the arrowed apple is picked up in scene A, there seems to be no undesired state in scene B. Likewise, the robot should anticipate the state of the surrounding objects (e.g., pink box) while fetching the green box. In the terminology mentioned earlier, the scene B appears to be static whereas the scene A is dynamic. Visual servoing is an attempt to tighly couple perception and action for handling the scene dynamicity, however, visual servoing remains reactive rather than anticipative or prospective. That is, a visual servoing-based system will detect that the cup spilled during pouring but will not anticipate it and therefore not prevent it. Notice that solving (b) contribute in solving (a) and vice-versa.

On the other hand, integrating robots in such human-centered environ- ments should also garantee safety and a step towards this goal is making the robots understanding what they are doing, in other words, building robots with explainable models (see Figures 1.9, 1.10). Though deep learning (DL) has shown great prowess on some perception tasks (e.g., object recognition, instance segmentation), there are more and more evidences that simply trying to compress huge amount of data especially when the data entropy becomes high fails to catch understanding: Slightly modifications of the pixels in images cause radical change in results or a DL telling that a train has been detected in the plate. This **(c)** **lack of explanability** in actual perception systems mostly relying on DL constitutes the third issue actual perception systems are struggling against.

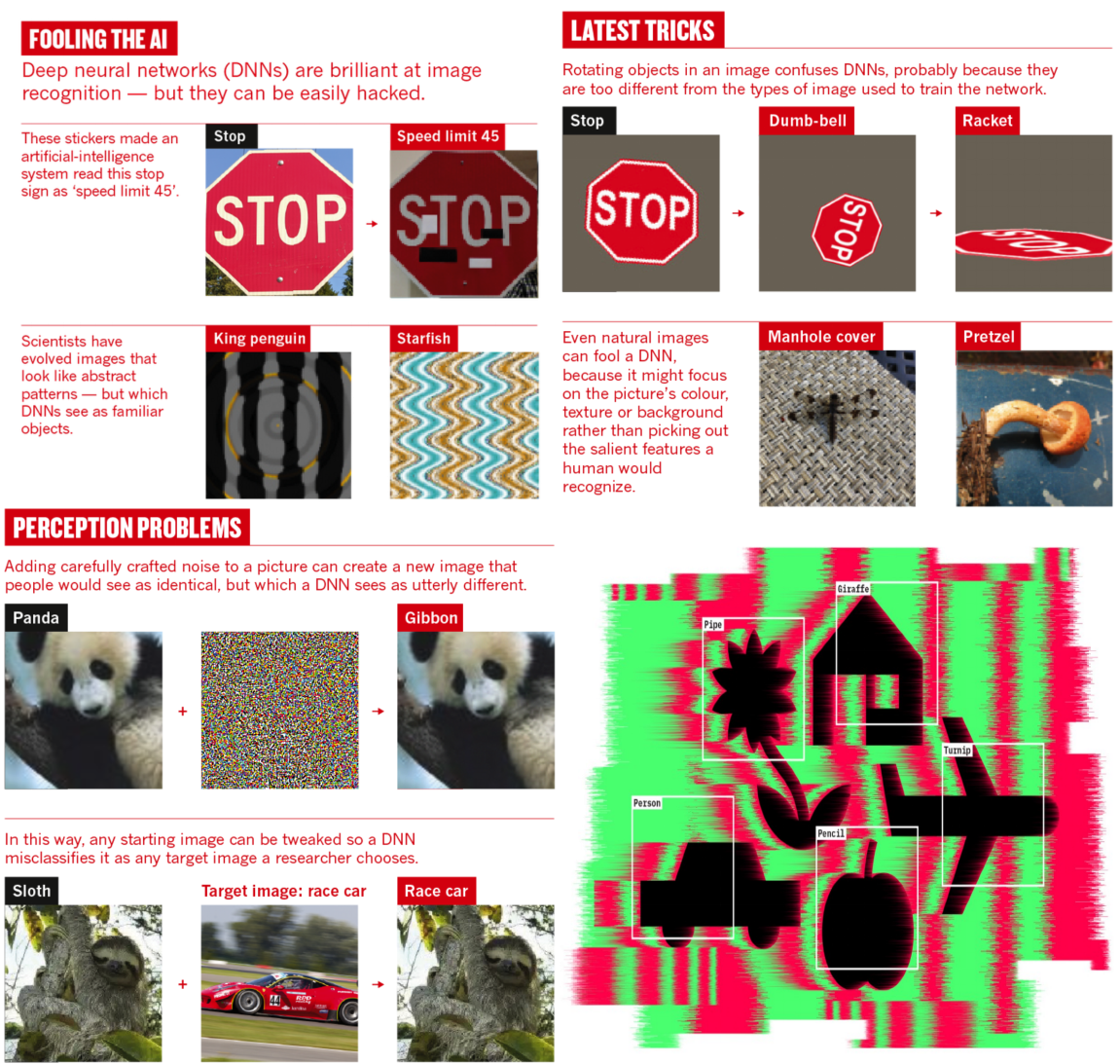


Fig. 2: Deep learning models are so easy to fool[[2]](#footnote-1)

Regarding the fooling of deep learning model mentioned above, GANs (Generative Adversarial Networks), is an attempt to fight against adversarial samples, which are fake samples resulting from sublte modifications of original samples. Unfortunately, though this algorithm is suitable for data integrity tests, it only worses the situation as far as perception is concerned, as it prevents robustness again variations at low level of semantics in the data. This is further presented by Figure 2 above.

1. **Research Problem**
   1. **Problem Statement**

Given the three issues (a-c) presented in the previous section, a question that arises is how do biological agents, at least humans, overcome them so trivially. In this regard, there are at least two observations. Firstly, **(i)** **Physics ultimately and constructively (i.e., in causal maner) deter- mines the world state**. Secondly, there are more and more evidences, in contrast to David Marr's view of perception, that **(ii) perception goes from the inside out, where a mental common intuitive physics engine continuously generate, simulate and maintains models of the world, which are then updated using sensory information** (see Figure 3)**.**

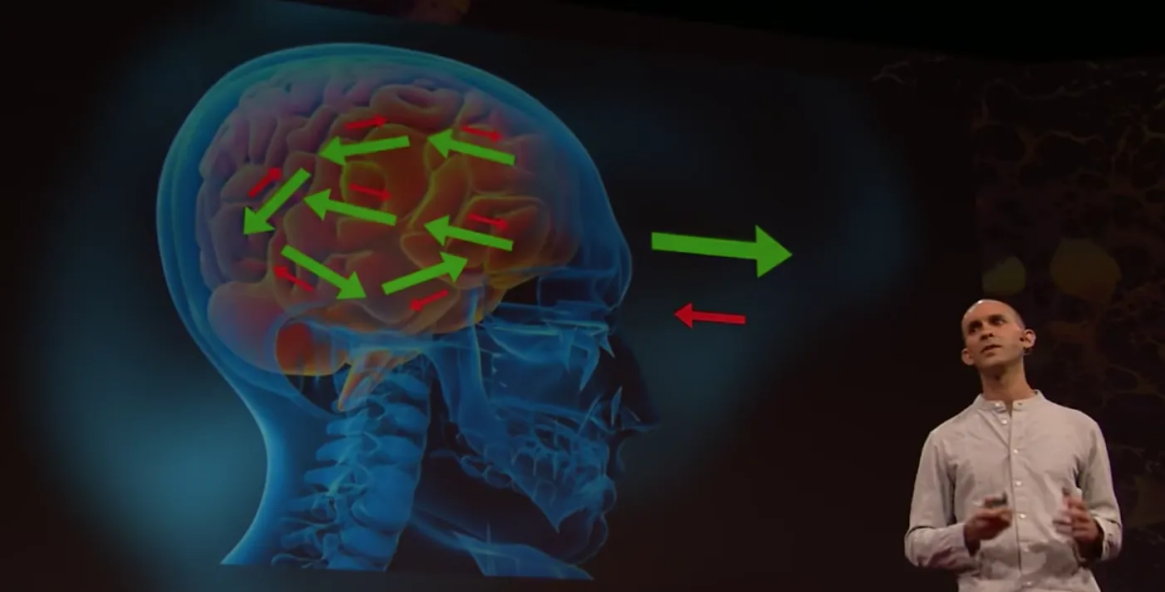


Fig. 3: Perception goes from the inside out

While the first observation **(i)** is a fundamental and supported premise of natural sciences, the second observation **(ii)** is corroborated by the first one and supported by recent research works summarized in the following three consistent points (**e)**:

**(e1) (Hesslow, 2022)** derived a theory of conscious thought as embodied mental simulation that sheds light on how anticipation takes place in biological agents, the importance of anticipation and the connection between sensory, motor, cognitive functions and the internal representa- tion of the world. Briefly, the agent mentally simulates its actions and perceives the resulting effects, then decides to proceed or not in the real world. And the loop restarts where the perception either mental or physical can then constitute the input to a new mental simulation.

**(e2) (Tenenbaum et al., 2020)** recently established five domains of human commonsense FPCIU, coined as the dark matters of perception, that invisibly drive our observational data and that artificial perception systems would have to meet in order to hope human-level perception, namely the Causality, Functionality, Intent, Physics and Utility. Note that before physics ultimately determines the world state, physics is also partially determined by the actions of agents, which are determined by their intentions, then utility and all these links are established by causal relations, which themselves are the basis for understanding and expplaining. On the other hand the world is supposed to host some activities that requires some facililities carrying some functions and knowing about such activities and functions inform about what constitutes the world state. This ability to catch from observation the physics that governs the behaviors of entities in the surrounding world and use it to anticipate the state of dynamic world is known as intuitive physics or physical reasoning. How this is actually implemented in the mind, there have been more and more evidences that physical reasoning takes place as mental simulations which are partial and probabilistic rather than pure reasoning with formal logical rules in order to cope with uncertainty and limited computational resources (e.g., realtimeness).

**(e3) (Anil Seth, 2018)** elaborated a theory of perception as controlled hallucination in order to explain how human perception overcomes the limitations of sensory information and copes with realtimeness, where the brain, called as bayesian, continuously guesses the world state (i.e., hallucination) and update it with the available sensory information (i.e., control).

Following this, the problem that this thesis tries to address is:

***(P)How can a robot perception system (NaivPhys4RP) be designed and implemented to leverage the physics that scene’s entities (e.g., manipulated objects, robots) as well as the agents’ sensory organs (i.e., sensors) undergo in order to anticipate and explain the state and observation of realistic worlds in an explainable manner with reasonable computational resources? NaivPhys4RP stands for Naive Physics for Robot Perception.***

* 1. **Problem Formalization**

In regard to the above theories, we formalize the problem addressed by NaivPhys4RP in four steps.

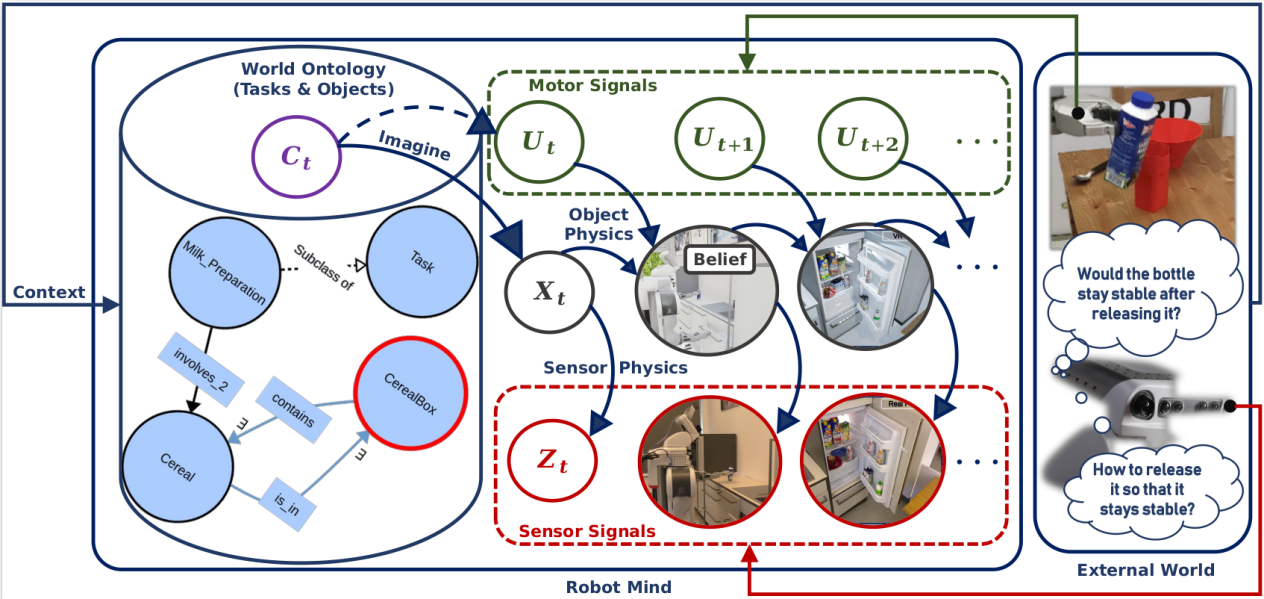


Fig. 4: Beyond static scenes, sensory information, and what-, where-ques- tions. Commonsense and especially intuitive physics, also coined as dark matter of perception, is a key for perception in dynamic and human-cente- red scenes. Perception as inner realistic world construction that anticipates and explains the world state as well as observations in an explainable manner, with reasonable computational resources

1. We model the world state, as shown by Figure 1, as a Situated (i.e., take place in a context) Partially-Observable (i.e., only partial sensor data) Hidden (i.e., not directly accessible information) Markov Process (i.e., state dependency) (SPOHMP) that evolves through the physics that scene entities (e.g., objects, robots, sensors) undergo. The context is supposed to catch other domains of the commonsense that drive the physics such as intentions, utility and functionality. Actions are already explicitly modeled. **(ii)** We model the hidden state a.k.a belief of the SPOHMP as a semantic digital twin of the world, cad a photo-realistic and physics-faithful replication of the world grounded in the world ontology for semantics. This makes the internal world representation suitable for emulating the SPOHMP. **(iii)** Then, we regard perception as taskable through queries and these perceptual queries are clustered into anticipatory (i.e., consequences given causes) and explanatory queries (i.e., causes given consequences), that are abstracted as the bayesian/ markovian inference tasks. However, note that an actual accurate and rich belief of the world state is the informational source for answering these questions. Such a belief is continuously filtered over time through emulation of the SPOHMP. **(iv)** Finally, we efficiently implement the four main operators of the rao-blackwellized particle filter, however modified to five operators, which is a generic, practical and constructive (i.e., explanability) approach to simultaneously emulate the SPOHMP and address the bayesian inference tasks just mentioned, through embodied, physics-faithful, photo-realistic, probabilistic, partial and ontology-grounded simulations. This formaliza- tion is summarized by the following system of equations (S1):

• X, is the world’s hidden state (i.e., a semantic digital twin of real world)

• Z, is the object/world observation (e.g., rgbd images)

• U, is the motion control (e.g., joint values, forces)

• C, is the process context (e.g., object + task knowledge)

Note that i, t, [.] and ∼ respectively denote the particle index, the time index, optional priors and the argmax probabilistic sampling.

1. **Research Goals**

The problem being formalized, the goal of this thesis regarding the design and implementation of NaivPhys4RP is on the one hand **(g2)** to efficiently compute the main operators of the particle filter (mRBPF) realizing the inference tasks in (S1). On the other hand and foremost, **(g1)** the goal will be to work towards a powerful representation of the world state that makes the computation of these inference tasks effective and efficient. In this way, this thesis solves the problem **(P)** defined in section 2.1.

**(g1) Belief state as semantic digital twin of the real world**

Such a representation (see Figure 5) goes beyond usual semantic scene graphs (objects’ description and relations among objects) and incorporates the scene geometry (e.g., articulated 3D models), scene physics (e.g., gravity, friction, mass, forces, viscosity, waves), scene agents (e.g., operating robots’ motorics and sensorics), scene ontology (i.e., semantics). For a possibly lossless representation and reliable simulation of the belief, the latter is directly represented in a photo-realistic and physics-faithful game engine, grounded in a rich scene ontology, and interfaces are provided to assert, modify, simulate and query it. In this way, the belief state is a replication of the real world (i.e., digital twin) in which every- thing is known and understandable (i.e., semantic).



Fig. 5: Belief (left), real world (right), world ontology (top).

**(g2) Effective and efficient computation of mRBPF’s main opera- tors**

• **(g2.1) Belief state initialization**

.The idea here is basically to sample the initial set of belief particlesbased on the initial context. The context variable circumscribes from the world ontology **(KB)** a set of concepts and relations among them, known as knowledge graph **,** that underlie the expected semantics of the scene. Then, as shown by Figure 5, this knowledge graph describing the context will be physically rendered as .This step is known as imagination and emulates the human ability to anticipate the world state given the context (e.g., activities, locations, intents) (see Figure 6).

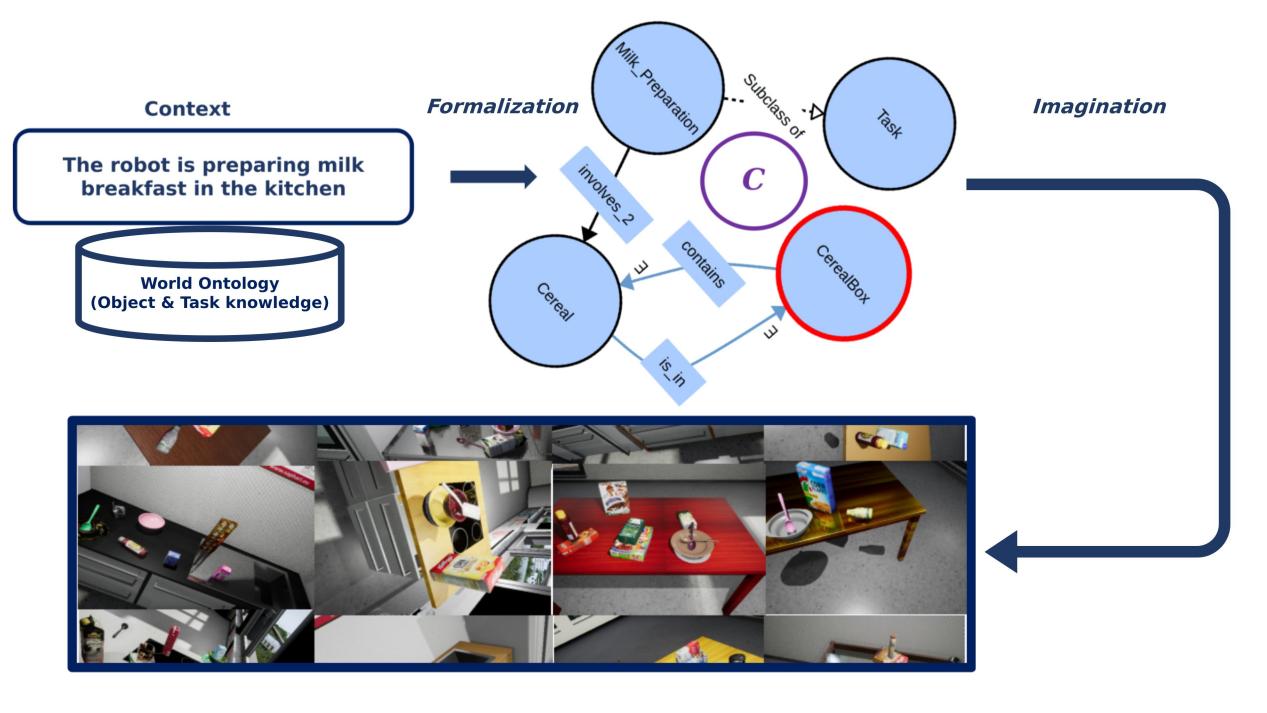
****

Fig. 6: Being aware of the robot activities and already imagining how the scene could look like.

• **(g2.2) Belief state prediction**

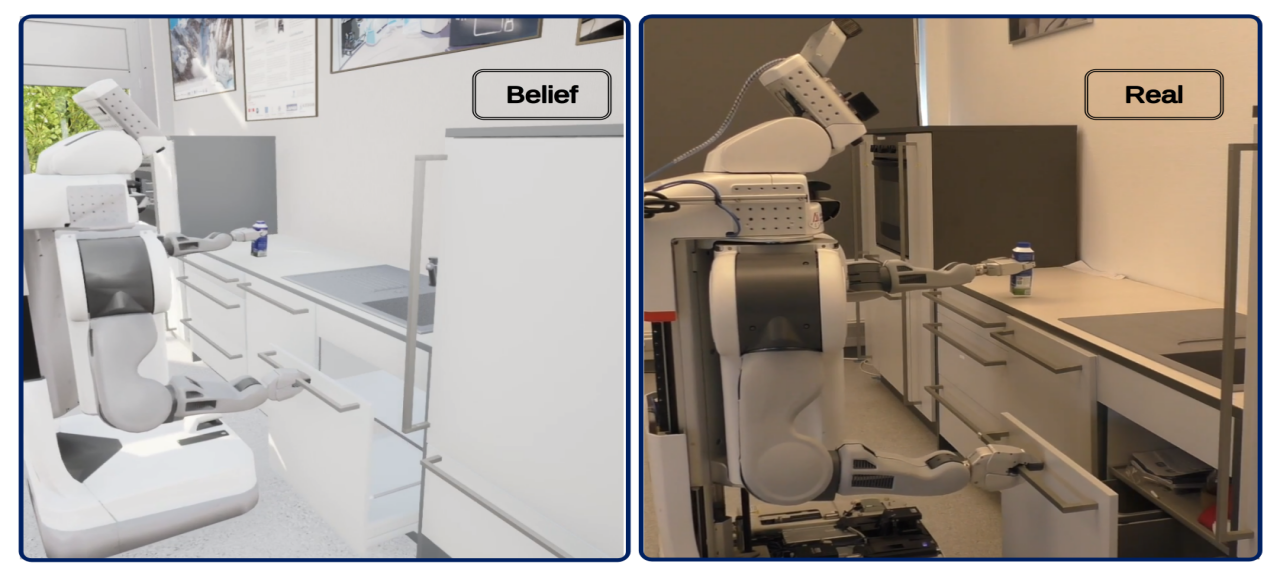


Fig.7: Forward embodied simulation for state anticipation

**.** This will be achieved by forward-simulating the motion of the operating agents in the actual world state (see Figu- re 7). Notice that more than being able to compute the different parallely, the sampling is a straight- forward operation, reducing therefore drastically the time complexity. Moreover, the realisticness of targeted in (g1) is crucial for the accuracy of the simulation.

• **(g2.3) Belief state augmentation**

**.** Note that the belief initialization in (g2.1) is only partial since the observation will be partial, the agent incrementally accesses the world state and a total initialization of the belief state is computationally complex. Then, forward simulating from such a partial initialization is not enough to achieve convergence of belief particles towards the world state. For this reason, a belief augmentation is performed after each prediction where identical operations as in the initialization step are used based on the actual context , and the results are then aggregated to the prediction for enriching it. At the belief initialization, there is no aggregation because the prediction is empty.

• **(g2.4) Belief state weighting**

**.** The weights of predicted beliefparticles

are computed by rendering the predicted state into and then computing the distance between and . Since, this weight correspond only to the observation **,** the cummulative weight that sums up all the weights of past observations is adopted:

**.** Notice that this operation is either straight- forward as it is a composition of straight-forward state rendering and straight-forward distance computation, reducing therefore the time com- plexity.

• **(g2.5) Belief state filtering**

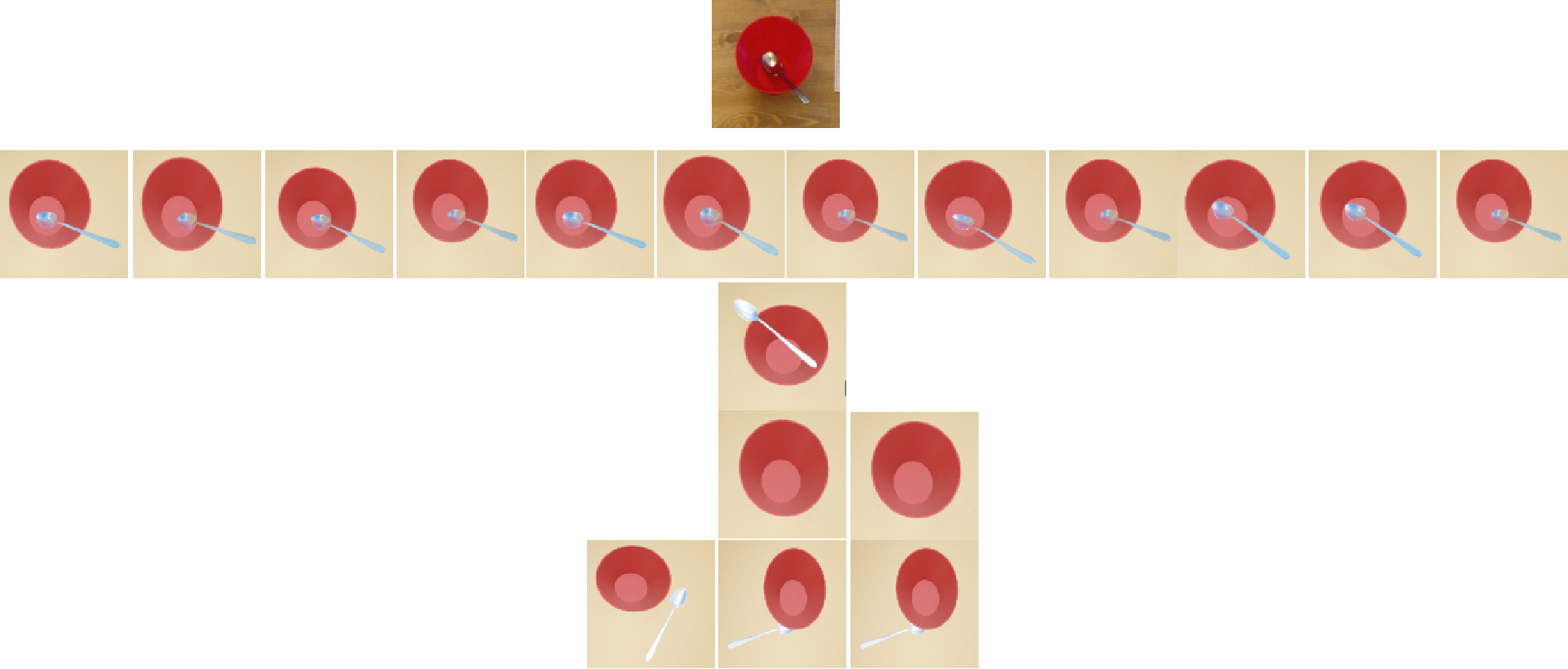
****

Fig. 7: Comparing & filtering rendered belief particles and real observation (top red-circled).

**.** The predicted particles are finally filtered based on their weights in order to ensure the converge of the set of belief particles towards the world state (see Figure 7). The complexity of this operation is linear to the number of particles.

• **(g2.6) World state explanation**

Though the native main operators of a mRBPF do not handle the explana- tion of states described as:

, their general princi- ples can be employed to address the problem. Literally, given the actual belief , we are looking for action in the context that would transform into a state within the context so that by applying the gi- ven action one could reach the given target state (e.g., how should I hold the milk bottle so that if I release it on the table, it will not fall). This problem is broken and solved using the general principles of operators in (g2.1-5) as follows:

**(g2.6.1) Target action initialization. :** The action is sampled given the actual context. When **U** is provided as evidences from the real robots, then it is enough and efficient to receive it as joint states that can basically be mapped onto the virtual robot to animate it. However, **U** as joint states are meaningless when sampling it. Instead, U will be sampled from a bag of generic primitive action plans then later contextuali- zed into complete motion plan therefore joint states based on and **.**

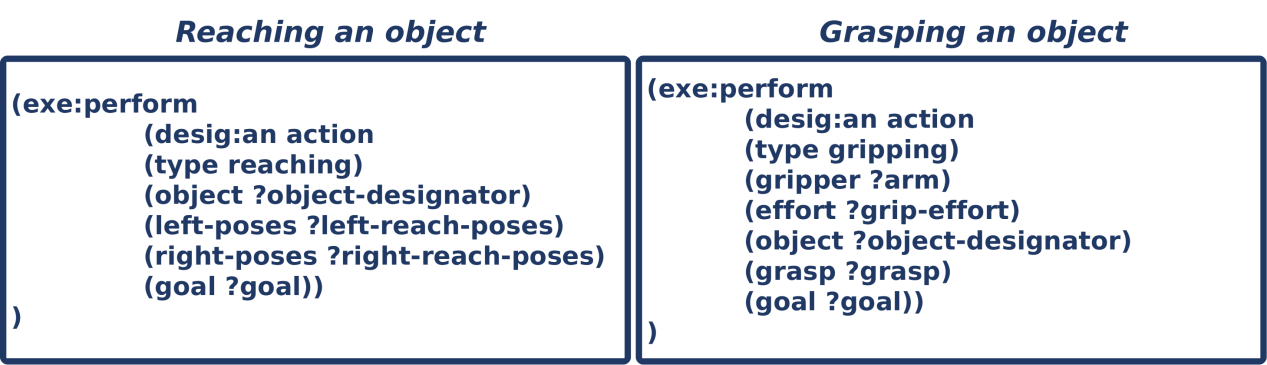
****

Fig. 8: Primitive action plans from CPL[[3]](#footnote-2) (CRAM Plan Language)

**(g2.6.2) Intermediate state prediction. :** Then, using the sampled action in (g2.6.1), this problem is solved using (g2.2) and (g2.3).

**(g2.6.3) Final state weighting. :** Then, using the sampled intermediate state in (g2.6.2), this problem is solved using (g2.6.2) while predicting and computing the weight as the distance between and the given desired state .

**(g2.6.4) Cause filtering. :** after sampling , the associated and are returned as the causes of the given desired state .

**Author’s Works**

1. ***Franklin Kenghagho Kenfack,*** *Michael Neumann, Patrick Mania, Toni Tan, Feroz Siddiky Ahmed, René Weller, Gabriel Zachmann and Michael Beetz, NaivPhys4RP - Towards Human-like Robot Perception: “Physical Reasoning based on Embodied Probabilistic Simulation”, In Humanoid Robots (Humanoids), 2022 (Submitted).*
2. *Patrick Mania,* ***Franklin Kenghagho Kenfack****, Michael Neumann, Michael Beetz, "Imagination-enabled Robot Perception", In International Conference on Intelligent Robots and Systems (IROS), 2021.* ***Best Paper Award on Cognitive Robotics****.*
3. *Gayane Kazhoyan, Simon Stelter,* ***Franklin Kenghagho Kenfack****, Sebastian Koralewski, Michael Beetz, "The Robot Household Marathon Experiment", In IEEE International Conference on Robotics and Automation (ICRA), 2021. Accepted for publication.*
4. ***Franklin Kenghagho Kenfack****, Feroz Ahmed Siddiky, Ferenc Balint-Benczedi, Michael Beetz, "RobotVQA --- A Scene-Graph- and Deep-Learning-based Visual Question Answering System for Robot Manipulation", In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, USA, 2020.*

**References**

1. *Hesslow Germund. (2002). Hesslow, G. Conscious thought as simula- tion of behaviour and perception. Trends Cogn. Sci. 6, 242-247. Trends in cognitive sciences. 6. 242-247. 10.1016/S1364-6613(02)01913-7.*
2. *Tenenbaum et al., Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense, In: Engineering, Volume 6, Issue 3, 2020, Pages 310-345, ISSN 2095-8099, <https://doi.org/10.1016/j.eng.2020.01.011.>*
3. *Anil K. Seth. “Consciousness: The last 50 years (and the next)”. In: Brain and Neuroscience Advances (2018).*

1. https://ease-crc.org/ [↑](#footnote-ref-0)
2. https://www.nature.com/articles/d41586-019-03013-5 [↑](#footnote-ref-1)
3. http://www.cram-system.org/ [↑](#footnote-ref-2)