2022 IEEE-RAS 21th International Conference on Humanoid Robots (Humanoids)
November 28-30, 2022. Ginowan, Japan.

NaivPhys4RP - Towards Human-like Robot Perception
"Physical Reasoning based on Embodied Probabilistic Simulation"

Franklin Kenghagho K.[1], Michael Neumann[1], Patrick Mania[1], Toni Tan[2],
Feroz Siddiky A.[1], René Weller[2], Gabriel Zachmann[2] and Michael Beetz[1]

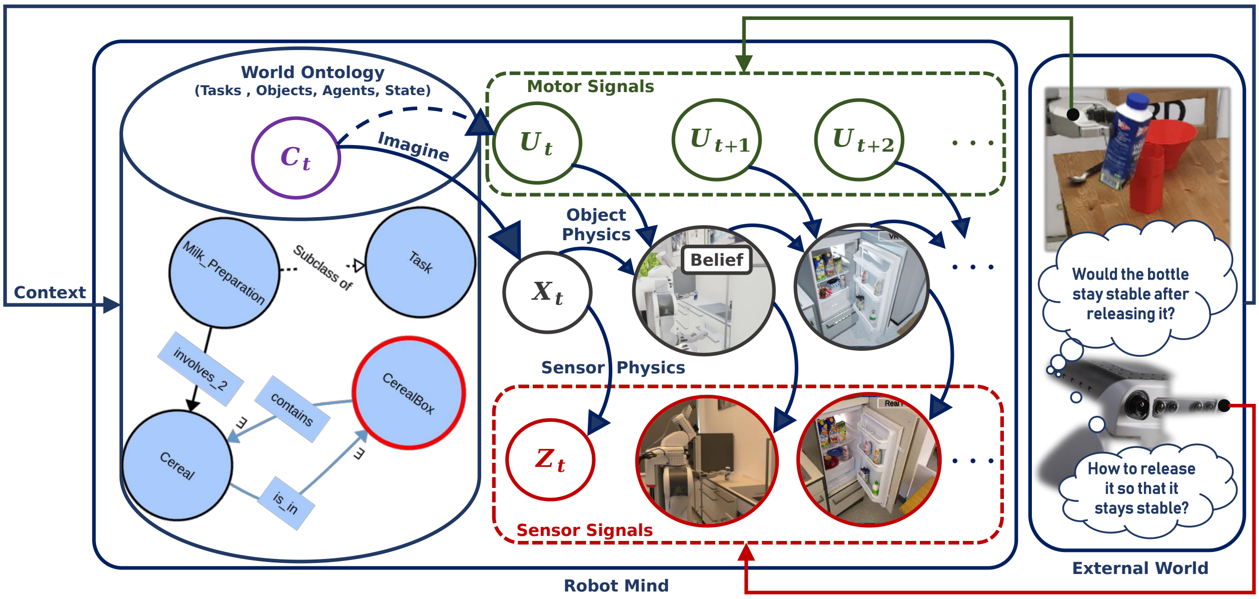


Fig. 1: Beyond static scenes, sensory information, and what-, where-questions. Commonsense and especially intuitive physics, also coined as dark matter of perception, is a key for perception in dynamic and human-centered scenes. Perception as inner realistic world construction that anticipates and explains the world state as well as observations in an explainable manner, with reasonable computational resources. We propose a white-box and causal generative model of perception in this paper.

Abstract— **Perception in complex environments especially dynamic and human-centered ones goes beyond classical tasks such as classification usually known as the what- and where-object-questions from sensor data, and poses at least three challenges that are missed by most and not properly addressed by some actual robot perception systems. Note that sensors are extrinsically (e.g., clutter, embodiedness-due noise, delayed processing) and intrinsically (e.g., depth of transparent objects) very limited, resulting in a lack of or high-entropy data, that can only be difficultly compressed during learning, difficultly explained or intensively processed during interpretation. (a) Therefore, the perception system should rather reason about the causes that produce such effects (how/why-happen-questions). (b) It should reason about the consequences (effects) of agent-object and object-object interactions in order to anticipate (what-happen-questions) the (e.g., undesired) world state and then enable successful action on time. (c) Finally, it should explain its outputs for safety (meta why/how-happen-questions). This paper introduces a novel white-box and causal generative model of robot perception (NaivPhys4RP) that emulates human perception by capturing the Big Five aspects (FPCIU)[1] of human commonsense, recently established, that invisibly (dark) drive our observational data and allow us to overcome the above problems. However, NaivPhys4RP particularly focuses on the aspect of physics, which ultimately and constructively determines the world state.**

I. INTRODUCTION

Manipulation/action in human-centered environments requires perception systems to inform about the state of the world. However, the actual perception systems are struggling against the extreme dynamicity of such environments as well as the safety required. On the one hand, **(a) sensor information are very limited**. With extrinsic and intrinsic limitations such as occlusion, delayed processing, missing or poor depth for smooth and glass objects, attempts to solely rely on these sensory information lead to a situation where compression while learning, interpretation and processing speed are no more efficient due to lack of or higher entropy in the data (e.g., hard pose estimation) [1]. On the other

*This work was not supported by any organization

[1] with Institute for Artificial Intelligence, Mathematics and Computer Science, University of Bremen, Germany fkenghag@uni-bremen.de

[2] is with Computer Graphics Research Group, Mathematics and Computer Science, University of Bremen, Germany toni@uni-bremen.de

[1] Functionality, Physics, Causality, Intention, Utility

979-8-3503-0978-2/22/$31.00 ©2022 IEEE

hand, **(b) these systems can only difficultly anticipate (undesired) states of the environment** given (a). Imagine the robot holding a plate containing a bowl and trying to open the drawer such as depicted by Figure 2.1, despite the fact that the robot camera is focused on the drawer, it 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 (see Figure 2.3).

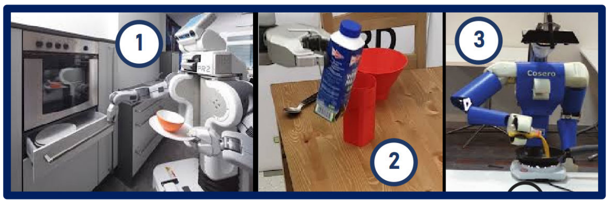


Fig. 2: Physical reasoning for perception in dynamic scenes

Notice that success depends on the robot's understanding of the milk's fluid dynamics and how to control it by manipulating the bottle in order to ensure that the milk will neither fall out of the mug, the mug will not spill, nor the mug will be overfilled (Frame 3). On frame 2, the robot should ensure that the blue milk will not fall after releasing it, which desired state does not only depend on the table's physical relief but also on some bottle's physical parameters such as the shape, volume, mass, content and height [10]. Visual servoing has been an attempt to catch this scene dynamicity, however it is not only just reactive rather than anticipative but not robust to sensory limitations mentioned above. Finally, **(c) robotics in human-centered environments should also ensure safety and a step towards this goal is making the robots understand what they are perceiving and doing, in order words our models should not only be explainable but explainable based on causality rather than associativity** unlike most recent developments on explainability [10]. Though Deep Learning (DL) has shown great prowess on some perceptual classification tasks, there are more and more evidence that simply trying to compress huge amount of data, especially when the data entropy becomes high, fail to catch understanding. Slight modifications of only few pixels in images cause radical changes in results or a DL-based model telling that a train has been detected in the plate [1]. Given these issues, we ask ourselves how biological agents, at least humans, overcome them. In this regard, there are at least two observations. Firstly, (1) Physics constructively and ultimately determines the world state. Secondly, (2) there are more and more evidences, in contrast to David Marr's view of perception, that perception mostly goes from the inside out, where a mental intuitive physics engine continuously generates, simulates and maintains models of the world, which are then updated using sensory information [10, 8, 4]. Such a perception theory is illustrated by Figure 1.

In this paper, we contribute in addressing the three issues mentioned above (a-c) by:

- proposing a **complete, practical, and modular** architecture of perception systems, coined as **NaivPhys4RP (Naive Physics for Robot Perception)**, that leverages the physics that manipulated scene objects as well as the agent's sensory organs undergo to anticipate and explain the state and observation of realistic worlds in an explainable manner with reasonable computational resources.
- providing a **proof of concept for NaivPhys4RP** by demonstrating it on different challenging scenarios, namely object-related (transparency, occlusion), task-related (i.e., pose estimation, stability check) and domain-related (kitchen, medical laboratory).
- Showing that **NaivPhys4RP substantially considers the Big Fives requirements FPCIU (Functionality, Physics, Causality, Intent, Utility)**[10] for achieving human-level perception recently established.

II. Related Work

Despite the increasingly intensive research on how biological agents, at least humans, do intuitively grasp the physical laws governing the state of the physical world around them from limited sensory information and how they apply such knowledge, commonly referred in the literature to as commonsense, intuitive, naive or folks physics, to anticipate the state or interpret observations, the results remain on the one hand abstract (e.g., higher-level hypotheses/findings) from the Psychology community [10] and primitive (e.g., 2D-, simplistic and unrealistic worlds, partial theories (e.g., disembodiedness)) from the community for computational sciences on the other hand [3] . This being said, we will mostly focus on the core computational theories underlying these research works as well as the two observations (1-2).

Embodied Simulation. Based on evidences, (Hesslow, 2002) [4] constructed a theory of conscious thought as embodied mental simulation, where the brain can simulate an action in an overt manner (i.e., without realization in real world) and simulate the perception of that action's effects usually referred to as Mind Eye, Ear, etc. Depending on the action's effects, the agent might decide to simulate the action in a covert manner, where the action is actually performed in the physical world. That action's effects are then perceived through the physical sensor organ (e.g., eye) and the cycle restarts. Note that, it is also possible to start the loop with a simulated perception from the mind eye (i.e., imagination). It is argued that the theory provides a way to the supportive interactions between motor, sensory, cognitive functions and the internal representations of the world, a way to anticipation a.k.a. prospection and emphasizes the essence of anticipation in cognition. (Cassimatis et al., 2004) emphasizes the advantages of the simulation theory of cognition and show how it constitutes a potential solution to many problems encountered in robotics.

Intuitive Physics. There have been more and more evidences that human cognition, yet at earlier months of life, can understand the physics governing the behavior of objects in the physical world and then use this knowledge to anticipate physical changes (i.e., object fall, object pose), which then enables successful and smooth action in realtime. Notice that this happens without prior education in physics or knowledge of the physical parameters of the world such as mass, friction, which are not only intractable and inexplicable for uneducated people in Physics but would not explain the smoothness and realtimeness of actions. In this regard, most research works have been supporting the hypothesis of a common physics engine that roughly infers the physical parameters (e.g., friction, mass) of the world from sensory information and then uses them as inputs to a forward simulation through the engine in order to anticipate events and states. Moreover, it has been shown that deviations in common physical reasoning could go back at least to the extrinsic (e.g., inaccurate physical parameters) and intrinsic (e.g., unobservable parameters) uncertainty of the physical phenomena, which parameters could be refined over time for more accurate reasoning. Researchers, especially Joshua Tenenbaum and his colleagues have considerably argued on how intuitive physics is essential for perception from limited sensory information (e.g., observing a car moving, and after it passes behind an occluding wall, we can still predict when it will appear at the other extremity of the wall) and have termed it as dark matter of perception in the sense that it is not directly graspable from sensory information but significantly contributes in generating these information [10]. However, Davis and his colleagues objected to the simulation theory for intuitive physics, claiming on the one hand the intractable computational resources required and on the other hand the failure of the simulation theory to the conjunction-fallacy effect. Recently, (Bass et al., 2022)[2] replied to Davis's objection with a theory of partial simulation. In sum, these works on intuitive physics stresses the physical, probabilistic, partial and emergent nature of the simulation theory of Hesslow.

Perception as Controlled Hallucination. (Anil Seth, 2018) [8] argues on the limitations of sensory information and flaws in David Marr's standard theory of perception (i.e., bottom-up information processing) and regarding this issue he elaborated a theory of perception based on evidences, where the brain, so-called bayesian, continuously generates, simulates expectations of the world state (i.e., hallucinations) and updates this expectations with the few available sensory information (i.e., control). This dominant top-down view of perception was already argued by (Ralf Moeller, 1996), defining perception as the process of anticipating sensory consequences of actions .

Imagination-capable Belief State (ICBS). Finally, we (Mania et al., 2021) [6] recently proposed a very rich inner representation of the world, also known as semantic digital twin as it aims at replicating the real world in photo-realistic and physics-faithful virtual environments (i.e., game engines) grounded in the world ontology for more semantics. Then, we showed how such a representation could be used to validate and refine the outputs of a traditional perception system. In this paper, we continue this work with regard of the above theories by enlarging the capabilities of these mental world representations to **embodied probabilistic** simulations and provide an architecture of perception systems, intrinsically based on such simulations and other aspects of commonsense such as the process context, that can perform physical reasoning to cope with the problems (a-c).

III. Architecture

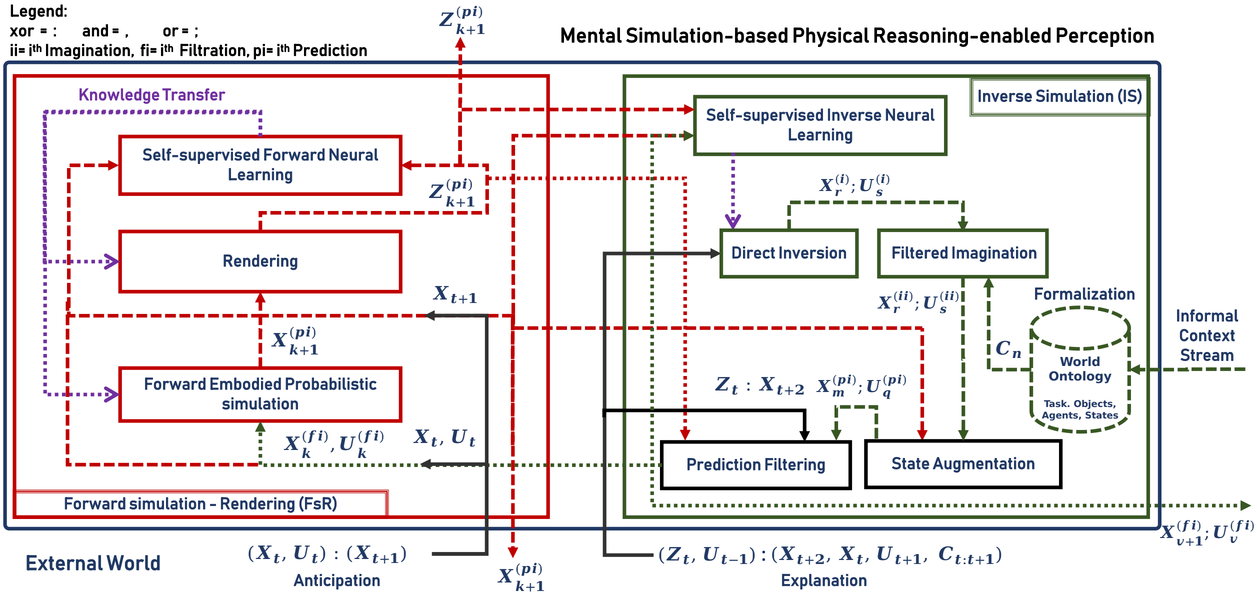

Legend:
xor = : and = , or = ;
ii= ith Imagination, fi= ith Filtration, pi= ith Prediction
$Z^{(pi)}_{k+1}$
Mental Simulation-based Physical Reasoning-enabled Perception
Knowledge Transfer
Self-supervised Forward Neural Learning
$Z^{(pi)}_{k+1}$
Rendering
X_{t+1}
$X^{(pi)}_{k+1}$
Forward Embodied Probabilistic simulation
$X^{(fi)}_{k}, U^{(fi)}_{k}$
X_t, U_t
Forward simulation - Rendering (FsR)
Inverse Simulation (IS)
Self-supervised Inverse Neural Learning
$X^{(i)}_{r}; U^{(i)}_{s}$
Direct Inversion
Filtered Imagination
$X^{(ii)}_{r}; U^{(ii)}_{s}$
Formalization
C_n
World Ontology
Task. Objects, Agents, States
Informal Context Stream
$Z_t : X_{t+2}$ $X^{(pi)}_{m}; U^{(pi)}_{q}$
Prediction Filtering
State Augmentation
$X^{(fi)}_{v+1}; U^{(fi)}_{v}$
External World
$(X_t, U_t) : (X_{t+1})$
Anticipation
$X^{(pi)}_{k+1}$
$(Z_t, U_{t-1}) : (X_{t+2}, X_t, U_{t+1}, C_{t:t+1})$
Explanation

Fig. 3: The robot observations Z_t and actions U_{t-1} are tightly coupled through a sufficiently rich inner model of the world X_t that allows through a forward simulation and rendering module (FsR) to anticipate the world state X_{t+1} ($X^{(pi)}_{t+1}$) and its observations Z_{t+1} ($Z^{(pi)}_{t+1}$), then to explain the world observation Z_{t+1} ($X^{(fi)}_{t+1}$) and its state X_{t+1} ($X^{(fi)}_{t}$, $U^{(fi)}_{t}$) through an inverse simulation module (IS). X_t emerges overtime through a complementary and white interaction loop between IS and FsR, where IS constructively infers the causes whose consequences through FsR match the observed or intended consequences.

A. Problem formalization

In regard to the above theories, we formalize the problem addressed by NaivPhys4RP in four steps. (i) We model the world state, as shown by Figure 1, as a **S**ituated (i.e., take place in a context) **P**artially-**O**bservable (i.e., only partial sensor data) **H**idden (i.e., not directly accessible information) **M**arkov **P**rocess (i.e., state dependency) (SPOHMP) that evolves through the physics that scene entities (e.g., objects, robots, sensors) undergo. (ii) We model the hidden state a.k.a. belief of the SPOHMP as ICBS described earlier. (iii) Then, we regard perception as taskable through queries [10, 5] 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 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 [7], however modified to five operators, which is a generic, practical and constructive approach to simultaneously emulate the SPOHMP and address the bayesian inference tasks just mentioned (markov-blanketed), through embodied, physics-faithful, photo-realistic, probabilistic, partial and ontology-grounded simulations. This formalization is summarized by the equations (1) below:

$$\begin{cases} \boldsymbol{X}_t^* \sim \boldsymbol{P}(\boldsymbol{X}_t|\boldsymbol{U}_{0:t-1},\boldsymbol{Z}_{0:t},\boldsymbol{C}_{0:t}) & \text{, actual belief} \\ \boldsymbol{X}_{t+1}^* \sim \boldsymbol{P}(\boldsymbol{X}_{t+1}|\boldsymbol{U}_t,\boldsymbol{X}_t,[\boldsymbol{C}_{t+1}]) & \text{, state anticipation} \\ \boldsymbol{X}_{t+1}^*,\boldsymbol{U}_t^* \sim \boldsymbol{P}(\boldsymbol{X}_{t+1},\boldsymbol{U}_t|\boldsymbol{U}_{t+1},\boldsymbol{C}_{t:t+1},\boldsymbol{X}_t,\boldsymbol{X}_{t+2}) & \text{, state explanation} \\ \boldsymbol{Z}_{t+1}^* \sim \boldsymbol{P}(\boldsymbol{Z}_{t+1}|\boldsymbol{X}_{t+1}) & \text{, observation anticipation} \\ \boldsymbol{X}_{t+1}^* \sim \boldsymbol{P}(\boldsymbol{X}_{t+1}|\boldsymbol{U}_t,\boldsymbol{X}_t,\boldsymbol{Z}_{t+1},\boldsymbol{C}_{t+1}) & \text{, observation explanation} \end{cases} \tag{1}$$

- X, is the world's hidden state (e.g., a digital twin)
- 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)

Following are the five main operators of the modified rao-blackwellized particle filter (*mRBFP*):

- Belief ***initialization***, $X_0^{(i)} \sim P(X_0|C_0)$
  amortized initialization, $X_0^{(i)} \sim P(X_0|C_0,Z_0)$
- Belief **prediction**, $\tilde{X}_{t+1}^{(i)} \sim P(\tilde{X}_{t+1}|X_t,U_t)$
- Belief ***augmentation***, $X_{t+1}^{(i)} \sim P(X_{t+1}|\tilde{X}_{t+1},C_{t+1})$
  amortized augmentation, $X_{t+1}^{(i)} \sim P(X_{t+1}|\tilde{X}_{t+1},C_{t+1},Z_{t+1})$
- Belief **weighting**, $W_{t+1}^{(i)} \approx P(Z_{t+1}|X_{t+1})$
- Belief **filtering**, $X_{t+1}^{(i)} \sim \frac{W_{t+1}^{(i)}}{\sum W}$

Note that i, t, $[.]$ and $\sim$ respectively denote the particle index, the time index, optional priors and the argmax probabilistic sampling. Though the variable U is not sampled by the above operators of a mRBPF, we show how the third equation in (1) can be solved using the general principles of these operators. Finally, the architecture on Figure 3 essentially computes these operators to solve the inference tasks in (1).

B. Ontology-Grounded Physico-Realistic Belief State (X_t)

An Imagination-Capable Belief State (ICBS) 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). The ontology is a formal description of fundamental and common truths about task-, agent-, object and state-related concepts, their properties and relationships among them in the scene. Depending on the particular scene under study, the ontology can be enriched with typical knowledge. It is also worth noting that state-related concepts that are unusual in most ontology definitions model in NaivPhys4RP a higher-level semantics of the effects of physics (e.g., through action) on the world. 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.

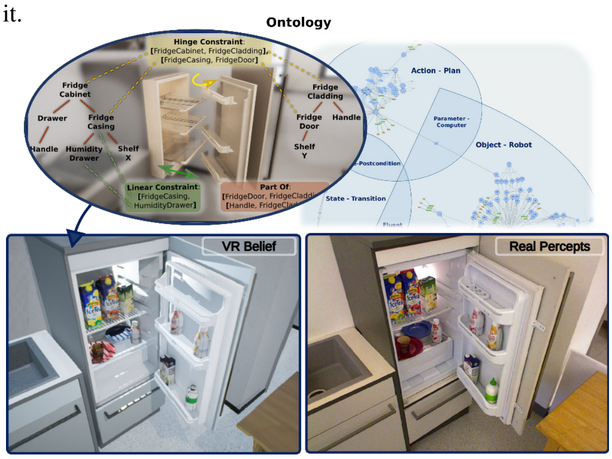


Fig. 4: Belief (left), real world (right), world ontology (top).

C. Forward Simulation - Rendering (FsR)

1) Anticipation: This FsR module, as reported in Figure 3's caption, is mainly responsible for anticipating the observations $Z_{t+1}^{(pi)}$ and the states $X_{t+1}^{(pi)}$ as consequences of the causes $X_t^{(fi)}$ and $U_t^{(fi)}$. Note that the superscripts pi and fi respectively denote the prediction p and the filtering f of particle i. Given our realistic mental simulations, these inference tasks are performed straight-forward as shown by the resolution equations (2) below:

$$\begin{cases} \boldsymbol{X}_{t+1}^{(pi)} \approx \boldsymbol{Simulation}_{\lambda_s}(\boldsymbol{X}_t^{(fi)},\boldsymbol{U}_t^{(fi)}) & \text{, state} \\ \boldsymbol{Z}_{t+1}^{(pi)} \approx \boldsymbol{Rendering}_{\lambda_r}(\boldsymbol{X}_{t+1}^{(pi)}) & \text{, observation} \end{cases} \tag{2}$$

The accuracy of these operations in (2) lies in the parameters λ_s and λ_r and we achieve it in two steps: targeting of realisticness (section III-C.2) and integration of uncertainty about physics (section III-C.3). For achieving a reasonable time complexity for the set of particles during inference, we rely as described below, on many cues such as parallelism, neural accelerators, Rao-Blackwellization and Partiality (section III-C.4).

2) Embodied Realistic Simulation: In the project **RobCog** (Robot Cognition: `robcog.org`) , as illustrated by Figure 4 and 5, we demonstrated how a photo-realistic and physics-faithful virtualization of everyday manipulation scenes (e.g., kitchens, medical labs) in the game engine Unreal Engine (UE) can be achieved, grounded in a large scene ontology (KnowRob-SOMA: `knowrob.org`) and used to perform human demonstrations of manipulation activities through a realistic human avatar so that rich datasets (NEEMs: Narrative-Enabled Episodic Memories) are automatically collected for machine learning purposes. The project **DAO** (Deep Action Observer)[2] extends RobCog by observing humans in activity and projecting their actions and motions onto programmable human avatars in the virtual world (see Figure 5). The project **URoboSim** (Unreal Robot Simulator: `embodied-ai.org/papers/URoboSim.pdf`), as illustrated by Figure 1 and 5, extends RobCog by developing virtual robot agents with sensing capabilities that can mirror what a real robot is doing or demonstrate what the real robot will be doing.

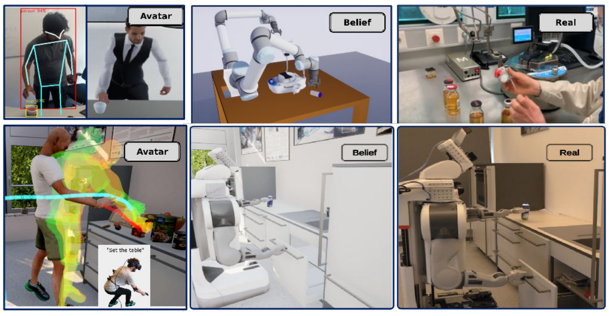


Fig. 5: RobCog (bottom-left), DAO (top-left), URoboSIM (real world in right and belief in left).

3) Uncertain Physics: Despite our ambition to target a realistic robot belief in appearance and physics, a perfect simulation remains challenging due to uncertainty about physical parameters like friction, mass, or object position in the world. In the belief X_t, uncertainty is partially considered in mRBPF as many belief particles are simulated, weighted, and then sampled based on their weights. However, this could require many belief particles to reach the right physical parameters, especially for continuous physical quantities. Collision and forces are fundamental in estimating the physical dynamics of objects in simulations. Therefore, to reduce the number of particles needed, we propose embedding uncertainty directly into object geometry, precisely, the underlying acceleration data structure. Within the scope of this paper, we applied the idea on top of *Inner Sphere Tree (IST)* [9]; nevertheless, it applies to other algorithms as well. As an example (see Figure 6), imagine the robot in Figure 5 trying to throw a blue milk bottle in the dustbin. In this case, the input is no more a single mass value of the object before its free-fall but rather a probabilistic distribution of its mass, friction, or object position. Likewise, the output will be a probabilistic distribution of its location when it finishes the fall. This approach considerably reduces the number of belief particles representing such distribution.

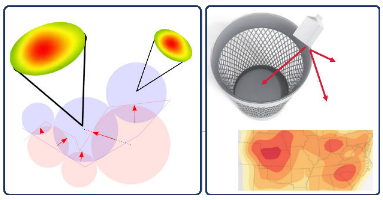
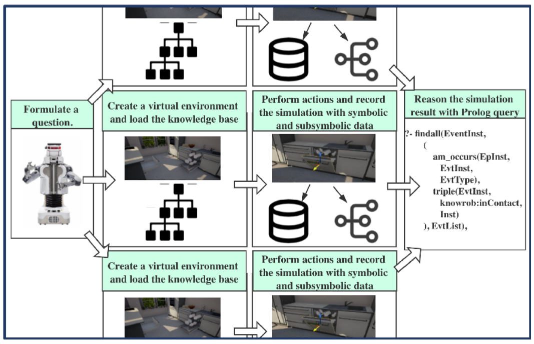


Fig. 6: (left) Elementary forces during a single simulation step between thrown bottle (blue) and dustbin (red), and (right) probabilistic distribution of bottle's location after simulation.

4) Temporal Efficiency: In this section, the cues we rely on to accelerate FsR on the set of belief particles are presented. **(i)** Rao-blackwellisation: Uncertain physical simulation is regarded as an emulation of the analytical estimation of probabilistic distributions of some continuous variables in X_t, reducing the number of belief particles needed for emulating the SPOHMP. **(ii)** Parallel FsR : We demonstrated in a Master thesis how, thank to cloud computing, FsR could be parallelized over the set of belief particles as shown by Figure 7. **(iii)** Partial simulation: SPOHMP is intrinsically partially-observable and this is taken into account during the emulation as X_t only get sampled incrementally through the augmentation operator of mRBPF. **(iv)** Self-trained neural accelerators: In [5], we demonstrated how a perception system can efficiently train on auto-generated data (e.g., NEEMs) from embodied and situated simulation to infer advanced semantic graphs of the scene. Instead of proceeding through procedural operator of the game engine, neural operators (λ_s and λ_r) trained from NEEMs could be integrated in game engines or operate beside them, as shown by the violet arrows on Figure 3. **(v)** Prediction as straightforward simulation: Finally, this is another major advantage of our approach over traditional symbolic and qualitative approaches which do not only require a huge gymnastics to sample from multidimensional probabilistic distributions, but also sample states that are not physically plausible within a certain context.


Formulate a question.
Create a virtual environment and load the knowledge base
Perform actions and record the simulation with symbolic and subsymbolic data
Reason the simulation result with Prolog query
Create a virtual environment and load the knowledge base
Perform actions and record the simulation with symbolic and subsymbolic data

```
?- findall(EventInst,
     (
       am_occurs(EpInst,
         EvtInst,
         EvtType),
       triple(EvtInst,
         knowrob:inContact,
         Inst)
     ), EvtList),
```

Fig. 7: Accelerating FsR through parallelism.

[2] dropbox.com/s/60fweieljn9pbky/deep-action- observer.pdf?dl=0

D. FsR-based Inverse Simulation (IS)

1) Explanation: This module is mainly responsible for processing explanatory questions such as presented in (1), in a constructive manner based on FsR, that makes it white and therefore interpretable and explainable, since FsR is eihter. Intuitively, the goal is to generate states $X_{t+1}^{(fi)}$ that explain observations Z_{t+1} and state-action couples $(X_t^{(fi)}, U_t^{(fi)})$ that explain desired states X_{t+1} and for achieving this, the remaining four main operators of mRBPF have to be computed.

2) (Amortized) Belief Initialization: It is intractable to merely sample these particles from the initial space of states. As humans rely on intuitive physics as a domain of commonsense to understand the physics that the world surrounding them undergoes, they do likely leverage commonsense about their operating scenes also referred to as context to formulate high-quality expectations about the scene state as far as the nature of objects and their natural (e.g., spatial) configurations are concerned in order to achieve estimation of the world state from limited sensory information. We model such a cognitive function in three core steps.

(i) Context formalization: As you can see from the architectural figures 3 and 1, the context that conceptually characterizes the scenes the robot operates in is either vaguely provided to the system under any communication modality such as text, audio, or even formally provided and directly stored within a shared memory. In the former case, The goal of the formalization step will then be to circumscribe a sufficiently rich field of concepts and relations among those concepts that underlie the target scene. Let assume that the most common input modality for context is textual, then our framework PRAC[3] (Probabilistic Action Cores) can be used to formalize such a vague specification, such as illustrated by Figure 8.

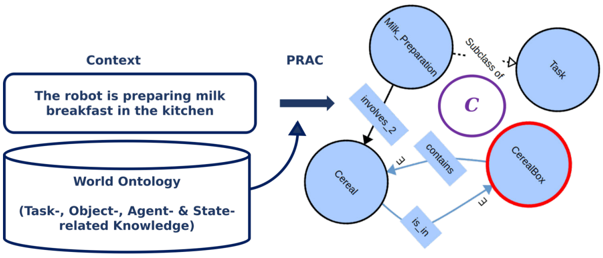

Context
The robot is preparing milk breakfast in the kitchen
World Ontology
(Task-, Object-, Agent- & State-related Knowledge)
PRAC
Milk_Preparation
Subclass of
Task
C
involves_2
Cereal
contains
∃
CerealBox
is_in
∃

Fig. 8: Context formalization.

(ii) Context-specific imagination: Once the context has been formalized, possible states of the world can be imagined.

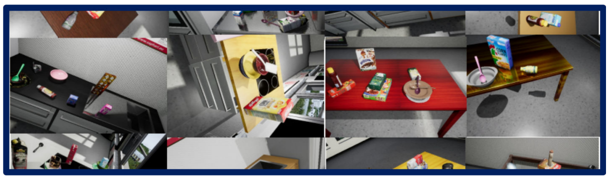


Fig. 9: Context-specific imagination of world state.

As shown by Figure 9, we demonstrated in [5] how situated and embodied datasets for perception systems could be generated from context-specific imagination. For preparing the breakfast, there is a need for cereal which is in the cereal box, a bowl and a spoon which can be and is usually inside the bowl.

(iii) Amortization: Despite the considerable reduction of the world state space through context-specific imagination, still there remains a bit of vagueness for instance in terms of number of objects and concrete spatial configurations. In order to amortize this combinatorial explosion, we employ a greedy direct (unconscious) perception approach of the scene, neurally trained on imagined datasets, to compress the state space. Then, the optimistic results of the neural learner are filtered based on the imagination (e.g., if knife detected then likely spoon because coffee drinking). We developed, RobotVQA (Robot Visual Question Answering) [5] for supporting the taskable and cognitive perception system RoboSherlock[4]. Notice that this step is realized by the *direct inversion* and *filtered imagination* modules on Figure 3.

3) (Amortized) Belief Augmentation: Notice that the belief initialization is only based on partial observations and the initialization is therefore only partial. 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 $X_{t+1}^{(pi)}$ where identical operations as in the initialization step are used based on the actual observation Z_{t+1} and context C_{t+1}, 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.

4) Belief Weighting: The weights of belief particles are $W_{t+1}^{(i)}$ computed by the straight-forward operation below:

$$\begin{cases} \boldsymbol{D}_{t+1}^{(i)} \approx \boldsymbol{Distance}_{\lambda_d}(\boldsymbol{Z}_{t+1}^{(pi)}, \boldsymbol{Z}_{t+1}) & \text{, actual} \\ \boldsymbol{W}_{t+1}^{(i)} \approx \boldsymbol{D}_{t+1}^{(i)} + \boldsymbol{W}_t^{(i)} & \text{, cumulative} \end{cases} \tag{3}$$

Intuitively, $D_{t+1}^{(i)}$ measures how close to the real partial observation Z_{t+1} the observation $Z_{t+1}^{(pi)}$ of the realistic rendering of the predicted belief $X_{t+1}^{(pi)}$ is (see Figure 4). For all the observations up to $t+1$ (i.e., total observations), the cumulative distance is expressed by $W_{t+1}^{(i)}$.

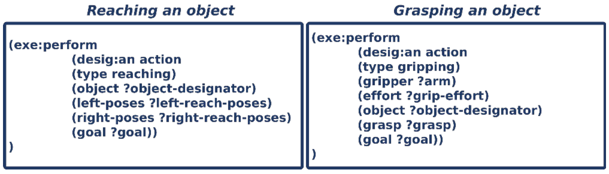
5) Belief Filtering: Finally, the belief particle are filtered through a random sampling with replacement according to their weights from the set of belief particles: $X_{t+1}^{(i)} \sim \frac{W_{t+1}^{(i)}}{\sum W}$. This ensures the convergences of the belief towards the real world state.

6) State Explanation: We highlighted earlier in this section that though the native main operators of a mRBPF do not support the explanation of states described as $\boldsymbol{X}_{t+1}^*, \boldsymbol{U}_t^* \sim \boldsymbol{P}(\boldsymbol{X}_{t+1}, \boldsymbol{U}_t | \boldsymbol{U}_{t+1}, \boldsymbol{C}_{t:t+1}, \boldsymbol{X}_t, \boldsymbol{X}_{t+2})$, their general principles can be employed to address the problem. Literally, given

[3] http://www.actioncores.org/

[4] http://robosherlock.org/, https://github.com/robosherlock

the actual belief X_t, we are looking for an action U_t^* within a context C_t that would transform X_t into a state X_{t+1} within a context C_{t+1} so that by applying the action U_{t+1} one could reach the target state X_{t+2} (e.g., how should I hold the milk bottle so that if I release it on the table, it will not fall). Notice foremost that this problem can be approximately broken into three problems according to rao-blackwellization namely (p1) $\boldsymbol{U_t^{(k)} \sim P(U_t|C_t)}$, (p2) $\boldsymbol{X_{t+1}^{(k)} \sim P(X_{t+1}|U_t, X_t, C_{t+1})}$ and (p3) $\boldsymbol{W^{(k)} \approx P(X_{t+2}|U_{t+1}, X_{t+1}, C_{t+1})}$. While (p2) and (p3) have already been solved by the FsR and Distance functions above, (p1) can be solved by sampling U_t according to the context C_t and the whole problem by filtering the U_t^* based on how good they turn X_t into the desired X_{t+2}. Notice the steps of a mRBPF except that U is the target instead of X. And since this work is about physical reasoning based on mental embodied simulations for perception, addressing the estimation of U to know about the state, does not only considerable goes beyond state estimation (e.g., action & motion planning required, U as joint states is not meaningful), but also emphasizes how perception, motorics and cognitive functions are strongly intertwined. In order to sample meaningful control commands U, we rely on CRAM (Cognitive Robot Abstract Machine)[5], an established cognitive architecture, that samples U from a bag of generic primitive action plans (see Figure 10), then contextualize it using the world ontology C and the world state X to finally produce joint states that can be directly realized by the virtual robots.



Reaching an object

```
(exe:perform
          (desig:an action
          (type reaching)
          (object ?object-designator)
          (left-poses ?left-reach-poses)
          (right-poses ?right-reach-poses)
          (goal ?goal))
)
```

Grasping an object

```
(exe:perform
          (desig:an action
          (type gripping)
          (gripper ?arm)
          (effort ?grip-effort)
          (object ?object-designator)
          (grasp ?grasp)
          (goal ?goal))
)
```

Fig. 10: Underspecified primitive action plans.

7) Temporal Efficiency: We leverage the following cues in order to achieve a reasonable time complexity for IS. **(i)** FsR's efficiency: IS is either a constructive approach based on FsR. **(ii)** Amortization: The use of self-trained neural accelerators for reducing the number of belief particles has been presented. **(iii)** Faster filtering: the belief particles are filtered based on a straight-forward computation of their importance weights. **(iv)** Faster convergence: The belief particles tend to converge quickly to the real world state since only few imaginary states are physically plausible before and after simulating.

IV. NaivPhys4RP and The Big Fives FPCIU

In a recent journal article [10], Tenenbaum and his colleagues identified five core aspects (FPCIU) of human commonsense, hierarchically organized, namely **F**unctionality, **P**hysics, **I**ntent, **C**ausality, and **U**tility to consider in order to hope human-level perception in Artificial Intelligence.

(i) Causality: As the basis for understanding, it is characterized by the elicitation of cause-effect relationships for the sake of explaining and anticipating phenomena. On the one hand, NaivPhys4RP inherently relies on physical simulation which itself relies on the integration of physical causality (i.e., laws of physics). Beyond physical causality, the context C encodes other forms of causality such as the functional causality (e.g., Milk preparation causes usage of certain products).

(ii) Physics: NaivPhys4RP obviously achieve commonsense physics through its ability to track the physical causality.

(iii) Functionality: Most objects in human-centered environments are functional and these functions are very decisive in experiencing (e.g., categorizing) the world around us though not directly observable from sensory information. NaivPhys4RP achieves this through functional causality.

(iv) Intent: in NaivPhys4RP, the agents's actions are modeled by the layer U even if in the current formalization, only the actions of the operating agent are explicitly represented. DAO can help in tracking and integrating other agents' actions. Moreover, the layer C (see Figure 8) partially captures the agents' intentions however can be made more explicit with an intent layer on top of U.

(v) Utility: humans act rationally by making choices that maximize their utility function (e.g., survival, travel cost, operation duration, success). NEEMs collected from NaivPhys4RP can be used for the learning of the agent's preferences such as scene objects, their spatial dispositions, the agent poses for grasping and perceiving different objects.

V. Experimentation

As a proof of concept, we demonstrate NaivPhys4RP in the following few challenging scenarios. We provide more information about the experiments in the demo video attached to this paper.

(i) 6D-Pose of In-hand Objects: Robots are usually unaware of the pose of objects they hold, which is critical for meaningful manipulation.

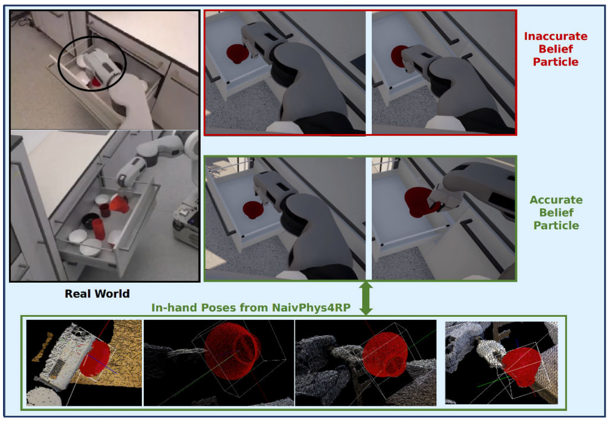


Fig. 11: NaivPhys4RP estimates in-hand poses.

(ii) 6D-Pose of Transparent & Smooth Objects: Certain objects exhibit a poor depth from optical depth cameras due to absorption, retransmission or specular rather than scattered reflection of emitted light rays. Figure 12 illustrates how

[5] http://www.cram-system.org/

NaivPhys4RP overcome the issue and estimate the pose of such objects.

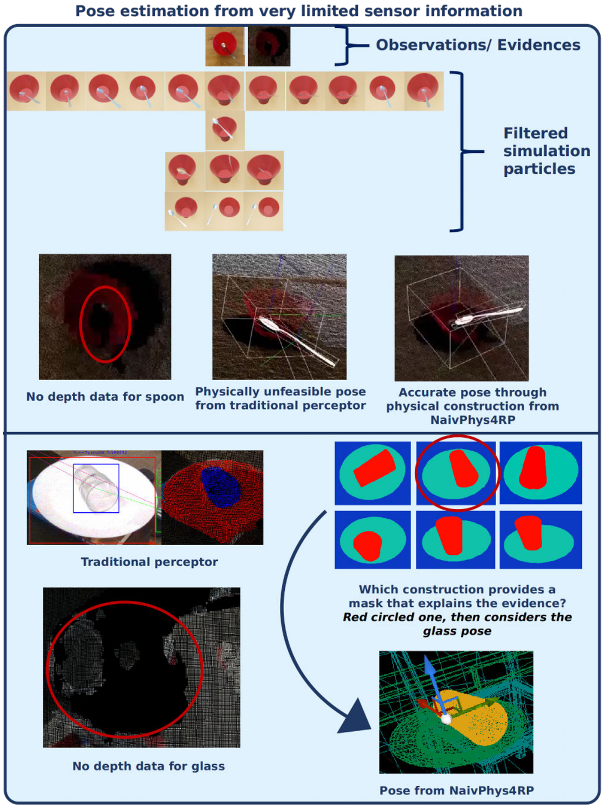

Pose estimation from very limited sensor information
Observations/ Evidences
Filtered simulation particles
No depth data for spoon
Physically unfeasible pose from traditional perceptor
Accurate pose through physical construction from NaivPhys4RP
Traditional perceptor
Which construction provides a mask that explains the evidence? Red circled one, then considers the glass pose
No depth data for glass
Pose from NaivPhys4RP

Fig. 12: NaivPhys4RP estimates poses from poor depth.

(iii) Object's Semantic Stability: How to place the milk bottle so that it does not fall?

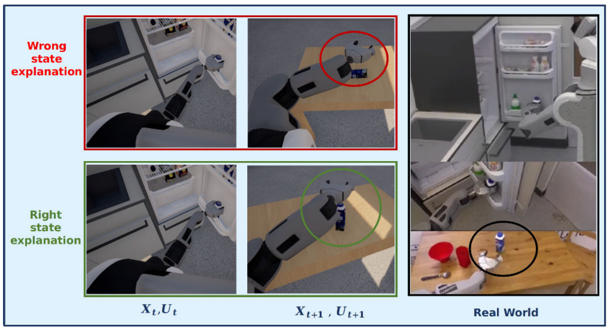


Fig. 13: NaivPhys4RP explains future desired state of world.

(iv) Generalizability: TraceBot:

Finally, we demonstrated how the approach is generalizable and can be applied to more complex, especially mission-critical applications such as TraceBot, a project that robotizes the process of medical sterility testing. Figure 14 shows how NaivPhys4RP can localize subtle tool parts and mirror the robot failures (`www.tracebot.eu`).

VI. CONCLUSIONS

We proposed in this paper a practical framework NaivPhys4RP with a proof of concept for scaling robot perception towards complex environments such as dynamic and human-centered scenes (i.e., motion, limited sensory information, safety). To emulate human perception, NaivPhys4RP essentially relies on realistic, embodied, physics-faithful and partial simulations grounded in the world ontology. In doing this, NaivPhys4RP substantially leverages commonsense knowledge about the world and foremost intuitive physics. In the future, we aim at a stable implementation of NaivPhys4RP with a focus on integrating the core components, but also on a systematic and quantitative evaluation and finally on an explicit integration of the FPCIU such as described in section IV.

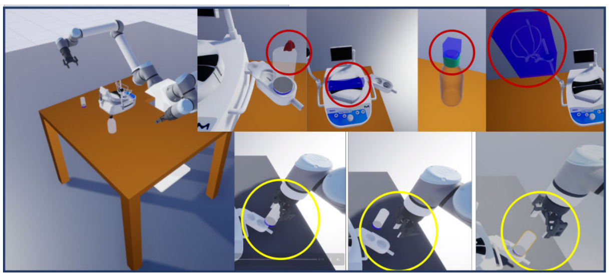


Fig. 14: NaivPhys4RP in TraceBot.

ACKNOWLEDGMENT

This scientific work is partially funded by the projects DFG EASE CRC 1320, EU TraceBot (grant agreement No 101017089), and BMBF ILIAS (grant no. 01DR19001B).

REFERENCES

[1] Pieter Adriaans. "Learning as Data Compression". In: 2007.

[2] Ilona Bass et al. "Partial mental simulation explains fallacies in physical reasoning". In: (2022).

[3] Jiafei Duan et al. *A Survey on Machine Learning Approaches for Modelling Intuitive Physics*. 2022.

[4] Germund Hesslow. "Hesslow, G. Conscious thought as simulation of behaviour and perception." In: *Trends in cognitive sciences* (2002).

[5] Franklin Kenghagho Kenfack et al. "RobotVQA — A Scene-Graph- and Deep-Learning-based Visual Question Answering System for Robot Manipulation". In: 2020.

[6] Patrick Mania et al. "Imagination-Enabled Robot Perception". In: 2021.

[7] Kevin Murphy and Stuart Russell. "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks". In: 2001.

[8] Anil K. Seth. "Consciousness: The last 50 years (and the next)". In: *Brain and Neuroscience Advances* (2018).

[9] Rene Weller and Gabriel Zachmann. "Inner sphere trees for proximity and penetration queries." In: 2009.

[10] Yixin Zhu et al. "Dark, Beyond Deep: A Paradigm Shift to Cognitive AI with Humanlike Common Sense". In: *Engineering* (2020).