A Formalisation of the Purpose Framework: the Autonomy-Alignment Problem in Open-Ended Learning Robots

Gianluca Baldassarre*†, Richard J. Duro*‡, Emilio Cartoni†, Mehdi Khamassi§,
Alejandro Romero‡, Vieri Giuliano Santucci†
*These two authors contributed equally to this work
†Institute of Cognitive Sciences and Technologies, National Research Council of Italy
{gianluca.baldassarre, emilio.cartoni, vieri.santucci}@istc.cnr.it
‡Integrated Group for Engineering Research, CITIC, Universidade da Coruña
{richard.duro, alejandro.romero.montero}@udc.es
§Institute of Intelligent Systems and Robotics, Sorbonne University / CNRS, Paris, F-75005, France
mehdi.khamassi@sorbonne-universite.fr

***Abstract*—[1] The unprecedented advancement of artificial intelligence enables the development of increasingly autonomous robots. These robots hold significant potential, particularly in moving beyond engineered factory settings to operate in the unstructured environments inhabited by humans. However, this possibility also generates a relevant *autonomy-alignment problem* to ensure that robots' autonomous learning processes still focus on acquiring knowledge relevant to accomplish human practical purposes, while their behaviour still aligns with their broader purposes (e.g., related to security and ethical constraints interference). The literature has only begun to address this problem, and a conceptual, terminological, and formal framework is still lacking. Here we address one of the most challenging instances of the problem: autonomous open-ended learning (OEL) robots, capable of cumulatively acquiring new skills and knowledge through direct interaction with the environment, guided by self-generated goals and intrinsic motivations. In particular, we propose a computational framework, first introduced qualitatively and then formalised, to support the design of OEL robot architectures that balance autonomy and control. The framework pivots on the novel concept of *purpose*. A human purpose specifies what humans (e.g., designers or users) want the robot to learn, do or not do, within a certain boundary of autonomy and independently of the domains in which it operates. The framework decomposes the autonomy-alignment problem into more tractable sub-problems: the *alignment* of 'robot purposes' with human purposes, either by hardwiring or through learning; the *arbitration* between multiple purposes; the *grounding* of purposes into specific domain-dependent *robot goals*; and the *competence acquisition* needed to accomplish these goals. The framework and its potential utility are further elucidated through the discussion of hypothetical example scenarios framed within it.**

Keywords: Formalised framework, open-ended learning, purpose, autonomy, arbitration, alignment, goals, grounding, competence acquisition,

[1]This work has received funding from: the European Union's Horizon 2020 Research and Innovation Programme, GA No 101070381, project 'PILLAR-Robots - Purposeful Intrinsically motivated Lifelong Learning Autonomous Robots'; the 'European Union, NextGenerationEU, PNRR', project 'EBRAINS-Italy - European Brain ReseArch INfrastructureS Italy', MUR code IR0000011, CUP B51E22000150006 and project 'FAIR - Future Artificial Intelligence Research', MUR code PE0000013, CUP B53C22003630006; and the European Innovation Council, GA No 101071178, project 'Counterfactual Assessment and Valuation for Awareness Architecture'.

I. Introduction

Current advances in artificial intelligence (AI) and robotics are yielding applications of significant value. These developments are largely driven by deep neural networks, the increased availability of data through widespread societal digitalisation, and the exponential growth of computational power [1]. This progress has spurred remarkable successes in fields such as computer vision [2], natural language processing and translation, and multimodal systems [3], [4]. Concurrently, AI advancements are enhancing the autonomous learning capabilities of robots, fostering a synergy between these fields [5]–[7].

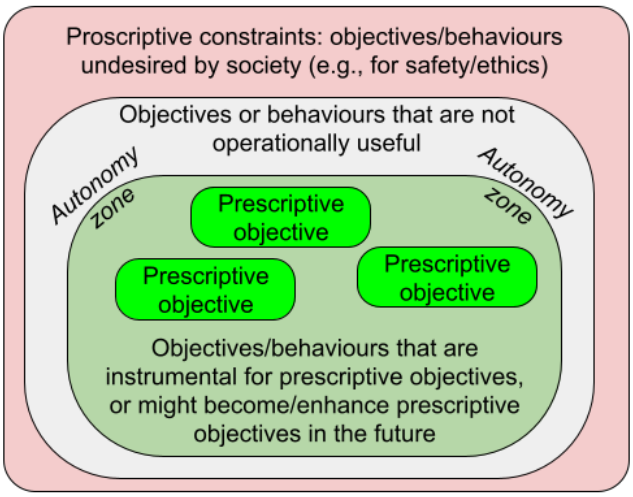

Fig. 1: The key elements of the autonomy-alignment problem.

This technological progress facilitates a significant transition: moving robots from predictable, engineered industrial settings to deployments within *unstructured, real-world environments inhabited by humans*, such as homes, offices, shops, and hospitals [8]–[10]. In these dynamic contexts, *autonomous learning* becomes crucial, enabling robots to acquire the knowledge needed to navigate challenges that are inherently *unpredictable at design time*. However, this

increasing autonomy simultaneously heightens the importance and complexity of the *AI alignment problem*.

Section VIII provides an overview of key topics within the AI alignment literature [11]. This review indicates that research predominantly addresses two main facets: *prescriptive aspects of alignment* –ensuring that AI systems pursue desired objectives and perform intended behaviors– and *proscriptive aspects of alignment* –preventing systems from exhibiting undesirable or harmful behaviors. This prescriptive/proscriptive distinction finds grounding in ethical philosophy concerning the nature of rules and norms [12]. While crucial, the existing literature offers comparatively less investigation into how the *autonomy* of robots can be effectively managed within these prescriptive and proscriptive boundaries, a challenge recognised in works on human-automation interaction and safe autonomy [13], [14].

We introduce the term *alignment-autonomy problem* to denote the set of challenges arising from the interplay between the need for robust alignment and the operational opportunities afforded by autonomy. This problem bears analogy to the classic *exploration-exploitation dilemma* in reinforcement learning [15], where an agent must balance exploiting known optimal behaviors against exploring potentially better, unknown alternatives. The alignment-autonomy problem, however, manifests at a higher level, concerning the *selection and pursuit of objectives* and overarching behavioral strategies, rather than just action selection within a fixed objective.

The alignment-autonomy problem requires trading-off multiple aspects (Figure 1). AI systems and robots must ensure *alignment* with human intentions and values, which involves adhering to both prescriptive goals and proscriptive constraints [16], [17]. Firstly, they should follow *prescriptive objectives* by actively pursuing objectives and performing behaviors desired by humans. Secondly, they must adhere to *proscriptive constraints* by avoiding objectives and behaviors deemed undesirable or harmful. Simultaneously, these systems should leverage the freedom afforded by their *autonomy* to best serve human interests. This includes pursuing objectives and performing behaviors that, while perhaps not explicitly prescribed or prohibited, are instrumentally beneficial for achieving overarching human goals. Importantly, for learning agents, this involves the capacity to autonomously discover novel objectives or acquire new skills that are instrumental to human-prescribed aims, or that serve an epistemic purpose by acquiring knowledge potentially useful for future goal achievement.

In this work, we address the alignment-autonomy problem by focusing on *open-ended learning (OEL) robots* [18]–[20]. The reasons for this focus are twofold. First, OEL robots represent a important class of autonomous robots, so addressing them covers a relevant portion of the overall problem. Second, OEL arguably presents the most difficult instance of the autonomy-alignment problem. Indeed, OEL robots *self-generate goals* under the drive of *intrinsic motivations* —-algorithms *purposefully designed to foster autonomous exploration and learning in the absence of human guidance* (e.g., demonstrations, externally assigned tasks, goals, or reward functions). By design, these robots have the highest propensity to explore and acquire behaviors that may misalign with human goals and values. Preventing such misalignment, without sacrificing the potential advantages of autonomy, poses a major challenge. Thus, knowledge gained by addressing the alignment of OEL robots can provide a valuable foundation for building frameworks and solutions applicable to other types of autonomous robots.

The core contribution of this work is the proposal of a *computational framework* for addressing the autonomy-alignment problem. The framework pivots on the novel concept of 'purpose'. *A human purpose specifies what humans (e.g., designers and users) want the robot to learn, do or not do, within a certain boundary of autonomy and independently of the domains in which it operates.* For example, a purpose of a designer may require that the robot, regardless of its deployment, must not harm people or damage objects. Another purpose, from a user, may want the robot to accomplish a specific operational goal, such as 'discard rotten fruit from the shop bench'. Yet another purpose, from another user, may require the robot to 'learn to manipulate fruit' for later assignment of more specific purposes.

The general idea is that designers and users can use purposes to specify particular tasks (prescriptive objectives), the generic *boundaries* within which robots should autonomously explore and acquire knowledge, or proscriptive constraints. To achieve this, the robot must encode each human purpose into an internal representation, the *robot purpose*. A key feature of the framework is that both human and robot purposes are *domain-independent*. This enables robots to pursue purposes across domains that are *a priori* unknown to them, and possibly even to their designers and users. Subsequently, within a given domain, robots can autonomously discover *domain-dependent robot goals* that fulfill the purposes, for example the purpose 'eliminate damaged fruits' could involve the acquisition of different goals depending on the types of fruits, containers, and contextual conditions.

The purpose-based framework also accommodates addressing *ethical issues*, specifically the need to prevent robots from behaving in ways that conflict with human values or social conventions, for example 'do not harm humans', 'do not break objects', 'do not interrupt people during conversations' [21], [22]. The domain independent nature of purposes might be useful in some cases. For example, the purpose 'do not cause harm to living beings' could protect animal species going beyond those known by the designer prescribing it. While the detailed treatment of these ethical aspects is out of reach for this work, we will outline how the purpose framework can be extended to incorporate such constraints, since purposes, and their *arbitration*, can specify outcomes and behaviors to be avoided.

The framework developed so far pivots mainly on *objectives*, understood as either abstract purposes or specific goals, each corresponding to particular states in the environment. This focus simplifies the broader reality that objectives may also involve more complex structures such as maintenance of states or ongoing processes, which are not considered here for simplicity and focus, and as end-state objectives are most common [23].

Overall, this work presents three novel contributions:

1) A conceptual and terminological framework identifying the fundamental elements for understanding and addressing the autonomy-alignment problem for OEL.
2) A formalisation of the framework and its concepts, paving the way for future mathematical analyses (theorem proving) and the development of specific robotic algorithms.
3) The use of the framework to decompose the broader alignment-autonomy problem into four specific, more tractable sub-problems involving purpose arbitration, human-robot alignment, purpose-goal grounding, and competence acquisition.

The remainder of the article is organised as follows. Section II reviews open-ended learning and the literature relevant for the autonomy-alignment problem. Section III introduces the concept of purpose and related notions (e.g., goals, alignment, grounding) in a qualitative form. Section IV presents the mathematical formalisation of the framework. Section V expands different possible types of purposes. Section VII analyses qualitatively an illustrative scenario through the framework to show it application. Section VI considers more in depth the four main sub-problems into which the autonomy-alignment problem can be decomposed. Section VIII overviews the main issues currently addressed in the literature on alignment. Finally, Section IX summarises the main contributions and outlines directions for future work. The *Appendix* illustrates the origin in cognitive sciences, of some concepts and terms employed in the framework.

II. Open-ended learning

A. Open-ended learning and limitations addressed here

In robotics and machine learning, OEL refers to a system's capacity to continuously acquire new knowledge and skills without predetermined tasks, enabling autonomous exploration and learning over time [18], [19], [24]. While machine learning approaches commonly train models on fixed datasets, OEL allows robots to acquire sensorimotor abilities in environments unknown at design time by progressively refining skills as new experience is gathered.

OEL shares similarities with *lifelong learning*, which also emphasises continuous knowledge acquisition, but with a higher focus on preventing catastrophic forgetting while new knowledge is acquired [25]. *Continual learning* is another related approach, where a system learns from a sequential data stream, although OEL remains more general by accommodating a wider variety of data sources [26]. *Curriculum learning* is another relevant method, focusing on externally structured task sequences to build capabilities. It differs from OEL as the agents commonly self-generate experience (hence the 'curriculum') based on their current lack of knowledge [27].

In OEL, the system is seen as maximising knowledge and skill acquisition, rather than optimising for task-specific rewards. A possible way to formalise this idea is to attempt to specify OEL objective function. One way to do this [24] is to assume that the robot explores the environment in a first 'intrinsic phase'. In a second 'extrinsic phase', the robot uses the acquired knowledge to maximise the performance across a set of externally assigned tasks unknown during the intrinsic phase:

$$\theta^* = \arg\max_{\theta} E_{g\sim\tau(g)} \left(E_{\pi(a|s,g,\theta)} R(g) \right) \tag{1}$$

where θ represents the robot controller parameters, g are goals 'drawn' from the environment, $R(g)$ is the reward function, and $\pi(a|s,g,\theta)$ denotes the goal-conditioned policy acquired by the robot during the intrinsic phase. The idea here is that the robot should be capable of autonomously acquiring knowledge and skills during the intrinsic phase to be ready to possibly solve *any* task in the environment that is assigned to it in the extrinsic phase. A strategy to support the robot's autonomous learning during the intrinsic phase is to employ *intrinsic motivations*, algorithms able to detect the acquisition of new knowledge and skills based on mechanisms such as novelty, surprise, competence improvement, mutual information, or empowerment [18], [28]–[31].

A notable limitation of OEL is that these autonomous learning processes are possibly 'too open' as they are agnostic to the actual purposes for which the user intends to employ the robots. So, one important risk is that the robots spend a lot of time and resources to acquire knowledge that is not useful for the users [32]. In addition, the robots may engage in behaviours diverging from user expectations, raising the need for frameworks that aligns autonomous learning with user purposes. The proposal of this work represents such a framework usable to develop OEL systems that, although still endowed with a remarkable degree of autonomy, tend to focus their learning processes towards the acquisition of knowledge and skills more aligned with the users' purposes.

III. Qualitative overview of the purpose framework

The purpose framework adopts terminology rooted in cognitive science (see Appendix). It is structured across three levels: the designer/user level, the robot level, and the domain level (Figure 2). The designer/user and robot levels each comprise two sub-levels. The designer/user and domain levels are external to the robot. Throughout the framework presentation, the term 'objective' is used neutrally, while more specific terms (e.g., robot purposes, robot goals, and domain goals) are introduced within each level to capture their distinct properties.

The first level concerns the designer/user, subdivided into the *human purpose* and *human goal* sub-levels. The human purpose sub-level encodes representations of objectives intended for robotic achievement in the environment. An example is 'sorting fruits into different containers'.

The human goal sub-level represents domain-specific instantiations of purposes. *Human goals* are internal representations of desired states in a given domain (level three). For example, human goals might specify 'bananas in a basket and pineapples in a crate' in one domain, and 'apples in a pot and pears on a plate' in another.

The second level concerns the robot, with sub-levels for *robot purposes* and *robot goals*. Robot purposes are domain-independent robot internal representations of human purposes.

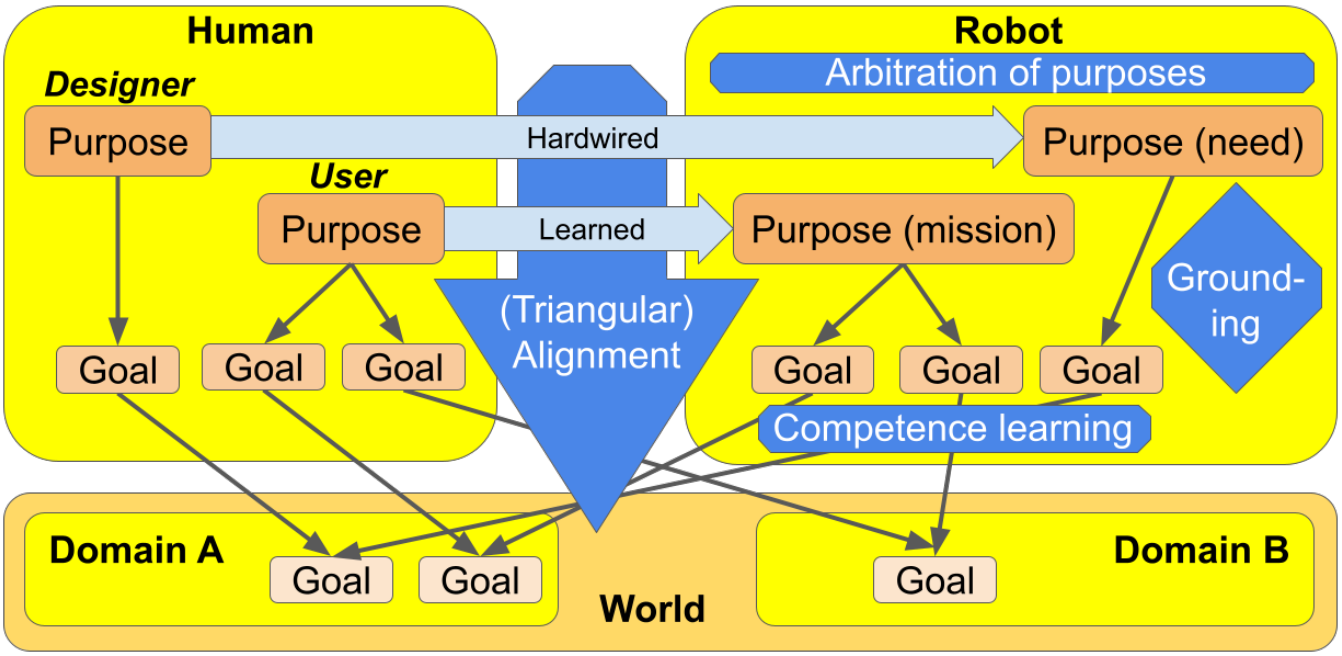

Fig. 2: **Main elements of the purpose framework.** The framework is structured into three levels. *Level 1* involves *humans* (acting as a *designer* or a *user*), who possess domain-independent *purposes* and domain-dependent *goals*. *Level 2* concerns the *robot*, endowed with domain-independent *purposes* –either hardwired (*needs*) or learned (*missions*)– and domain-dependent *goals*. *Level 3* comprises the *domains*, each characterised by *state-goals* corresponding to robot and human goals. A *triangular alignment* (*alignment* for short) occurs when a human purpose and its corresponding human goal, and a robot purpose and its corresponding robot goal, converge on the same state goal, indicating coherent alignment between human and robot objectives.

This abstraction enables generalisation across domains and underpins the robot's learning processes including the acquisition of goals, skills, and world models.

Robot purposes can be *hardwired* by designers, in which case they are termed *needs*. Needs mirror phylogenetic motivations in biological systems. Examples include a *homeostatic need* such as 'maintain battery charge' or an *epistemic need* such as 'acquire fruit images' for training internal classifiers. In general, needs can be seen as designer-encoded purposes.

Alternatively, robot purposes can be autonomously acquired by the robot through learning, aiming to align with human purposes. Such purposes are termed *missions*. An example of mission is 'sort fruits into different containers.' The acquisition of missions can be guided by internal criteria derived from designer purposes. Such guidance may be implemented as hardwired needs or encoded within the robot's learning architecture, for example, mechanisms promoting interaction with users to infer and internalise their purposes.

The robot-goal sub-level encodes *robot goals*: observation-based representations of robot purposes instantiated within specific domains. Robot goals specify desired domain states.

The third level is the *domain level*, comprising the robot's external physical/social environment and its sensorimotor body. Each robot goal corresponds to a specific domain state termed *state-goal*. Similarly, human goals correspond to state goals across domains. When a robot purpose aligns with a human purpose, both yield corresponding state goals across domains, a condition termed *triangular alignment* (or simply 'alignment').

Figure 3 schematically illustrates the framework. At the human level, purposes are domain-independent, with possible domain-specific user-goals. At the robot level, robots possess a *motivational space* formed by robot purposes corresponding to human purposes. This space may include: a learned mission (e.g., 'sort fruits into containers'), an epistemic need (e.g., 'learn to manipulate fruit'), and a homeostatic need (e.g., 'maintain battery charge'). For instance, robot 1 seeks fruits and is curious, while robot 2 seeks fruits and energy maintenance. Each robot purpose is depicted along one dimension, although purposes are typically multidimensional. Robot purposes are associated with utility functions (blue-to-red gradients), potentially peaking at an ideal set-point (marked by a smiley face). Robots also possess an observation space encoding domain-specific robot goals linked to the related purposes. At the domain level, multiple state-goals can satisfy the same mission. Examples include filling containers with pears (state-goal 1.1), apples (state-goal 1.2), bananas (state-goals 2.1 and 2.2), or pineapples (state-goal 2.3).

The framework allows to decompose the autonomy-alignment problem in relevant sub-problems:

- *Human-robot alignment*: how to ensure that robot needs or autonomously learned missions are aligned with human purposes.
- *Purpose arbitration*: how to prioritise among multiple concurrent purposes.
- *Purpose-goal grounding*: how to enable robots to acquire domain-specific goals that optimally fulfill purposes;
- *Competence acquisition*: How to ensure that robots ac-

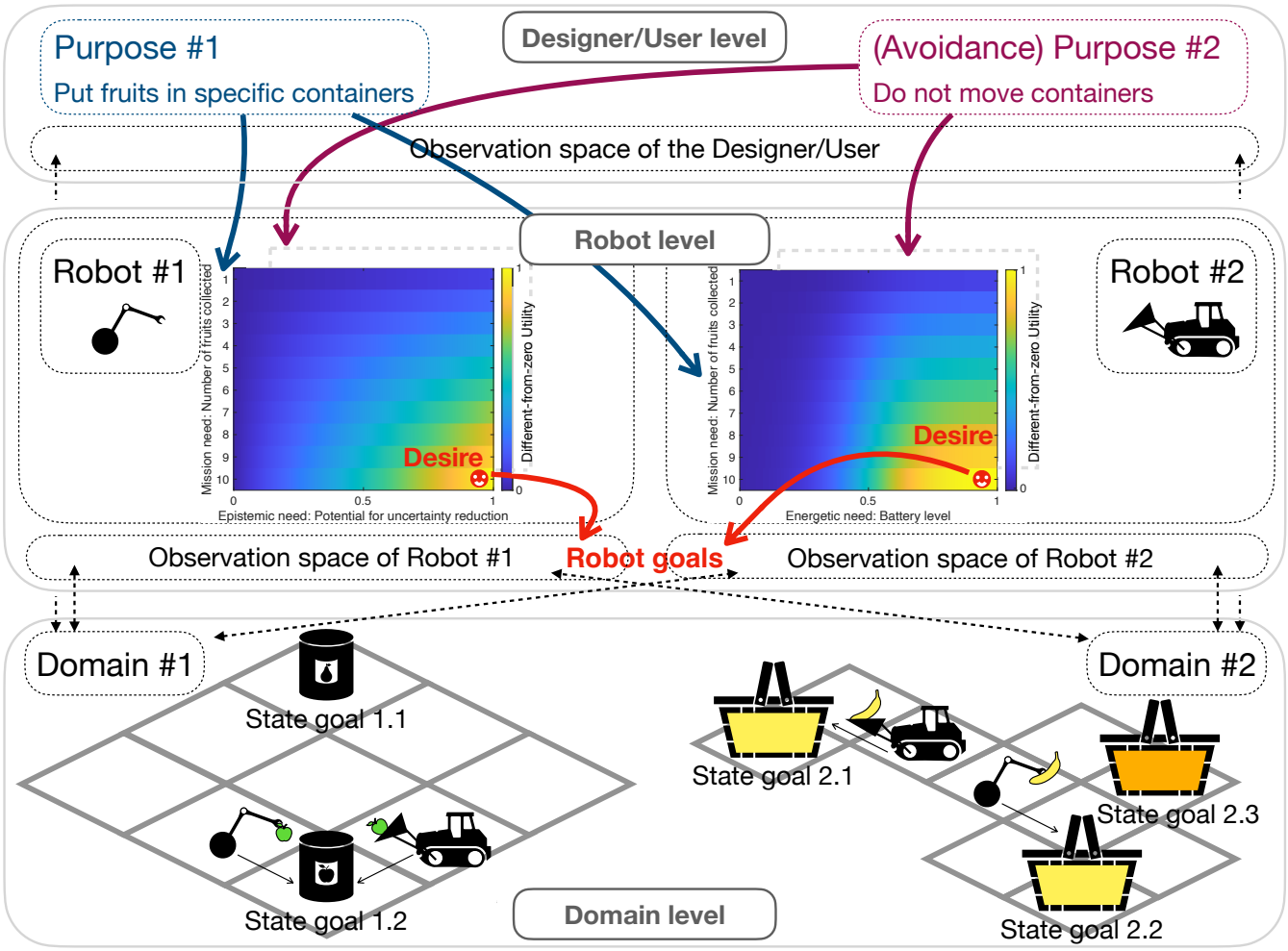

Fig. 3: **Illustrative example of the purpose framework.** See text for details.

quire the skills needed to accomplish the goals.

IV. Formalisation of the Purpose Framework

This section formalises the purpose framework and its core constructs. Figure 4 summarises the main elements. The formalism adopts the perspective of an external observer (e.g., a researcher) observing designers and users of robots, the robot controller, and the world comprising different domains (here, domains include both the robot's sensorimotor body and its external environment).

We frame the formalisation from a *goal-based perspective* [33], [34], grounded in the formalism of Markov Decision Processes (MDPs) used in reinforcement learning (RL) [15].

a) Notation: Lowercase letters denote elements of sets, capital letters denote sets. Subscripts indicate indexing (e.g., O_c is robot c's observation set). Superscripts specify symbols (e.g., U^E is the utility over an encoding space E). Functions are denoted with f, with superscripts indicating their domain and codomain (e.g., f^{O-P} maps observations to purposes). Given a discrete set S, $\Delta(S)$ denotes the probability simplex over S. Sets are sometimes referred to as *spaces* to highlight internal structure (e.g., an observation space O with similarity relations).

Table I summarises the symbols used in the formalisation.

b) Core elements: Different *humans* (designers/users) are indexed by $h \in H$; different *robots* (or 'cobots' –collabroative robots) by $c \in C$; different *domains* by $d \in D$.

c) Domains: Time is discretised as $t \in \{0, 1, 2, ...\}$. Each domain $d \in D$ is characterised by states $s_d \in S_d$, where S_d is the *domain state space*. The domain transition function is:

$$f_{d,c}^{SA-S} : S_d \times A_c \to \Delta(S_d),$$

defining the probability of transitioning from $s_{d,t}$ to $s_{d,t+1}$ under action $a_{c,t} \in A_c$ chosen by robot c.

d) Human encoding spaces and purposes: Each human h possesses multiple *encoding spaces* $E_{h,i}$, indexed by i, comprising points $e_{h,i} \in E_{h,i}$.

A *purpose* $P_{h,i} \subset E_{h,i}$ is defined as:

$$P_{h,i} = \{e_{h,i} \in E_{h,i} \mid f_{h,i}^{E-U}(e_{h,i}) \neq 0\},$$

where $f_{h,i}^{E-U} : E_{h,i} \to U_{h,i}^E \subseteq \mathbb{R}$ is the *purpose utility function*. Humans may have multiple purposes $P_{h,i} \in \mathcal{P}_h$. *Avoidance purposes*, based on negative utilities, can be defined to encode undesired outcomes, critical for safety and ethical alignment.

The *human observation-encoding function* is:

$$f_{h,i}^{O-E} : O_h \to E_{h,i}.$$

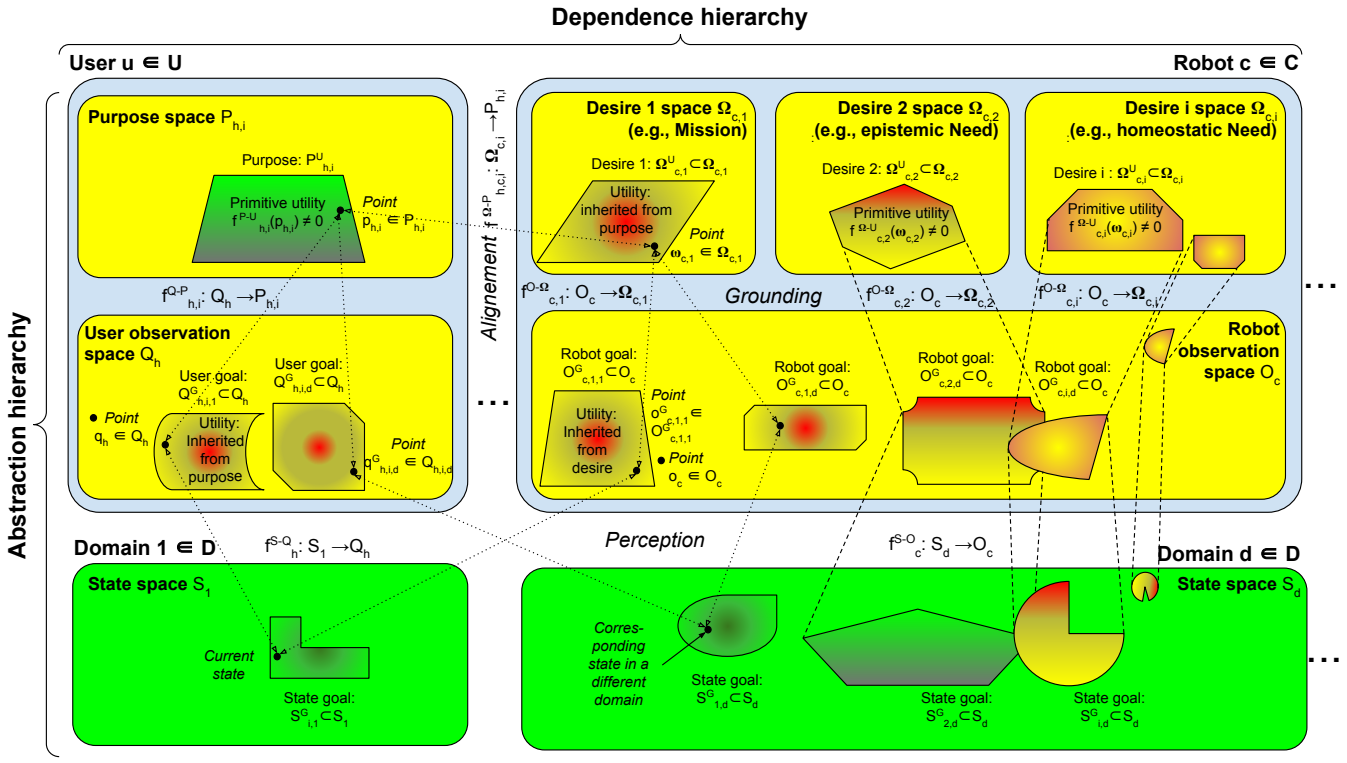

Fig. 4: **Main elements of the purpose framework and associated formalism.** Multiple domains are considered (here two are shown, in green). The user hosts several purpose spaces (one shown, in yellow), each abstracting observations and including a purpose, with a utility gradient over its elements. Each purpose corresponds to distinct *user-goals* across domains, themselves subsets of observations, inheriting utility from the purpose. User-goals map to sets of states (*state-goals*) in the domains. Multiple robots may serve user purposes (one shown). The robot hosts multiple purpose spaces, including *missions* (learned purposes) and *needs* (hardwired purposes). Robot purposes correspond to *robot-goals* across domains. Dotted lines illustrate that for effective service, robot-goals should align with user-goals at the level of state-goals.

e) Human observation space and goals: Each human h has an *observation space* O_h, populated via:

$$f_h^{S-O} : S_d \to \Delta(O_h).$$

Each purpose $P_{h,i}$ induces a different *human goal* in each domain d:

$$G_{h,i,d} = \{o_h \in O_h \mid f_{h,i}^{O-E}(o_h) \in P_{h,i}\}.$$

Each observation inherits the purpose utility:

$$f_{h,i}^{O-U}(g_{h,i}) = f_{h,i}^{E-U}(f_{h,i}^{O-E}(g_{h,i})).$$

Each human goal corresponds to a *state-goal*:

$$G_{i,d} = \{s_d \in S_d \mid f_h^{S-O}(s_d) \in G_{h,i,d}\}.$$

f) Robot encoding spaces and purposes: Each robot c has multiple *encoding spaces* $E_{c,i}$, with $e_{c,i} \in E_{c,i}$.

A *robot purpose* $P_{c,i} \subset E_{c,i}$ can be either a hardwired need $N_{c,i}$, or a learned mission $M_{c,i}$.

Robot needs are defined based on a hand-crafted utility function:

$$N_{c,i} = \{e_{c,i} \in E_{c,i} \mid f_{c,i}^{E-U}(e_{c,i}) \neq 0\},$$

where $f_{c,i}^{E-U} : E_{c,i} \to U_{c,i}^E \subseteq \mathbb{R}$.

Robot missions are defined through an *alignment function*:

$$f_{h,c,i}^{E-E} : E_{c,i} \to E_{h,i},$$

such that:

$$M_{c,i} = \{m_{c,i} \in E_{c,i} \mid f_{h,c,i}^{E-E}(m_{c,i}) \in P_{h,i}\}.$$

Utilities propagate from human ones as:

$$f_{c,i}^{E-U}(m_{c,i}) = f_{h,i}^{E-U}(f_{h,c,i}^{E-E}(m_{c,i})).$$

The *robot observation-encoding function* is:

$$f_{c,i}^{O-E} : O_c \to E_{c,i}.$$

g) Robot observation space and goals: Each robot c has an *observation space* O_c, populated via:

$$f_c^{S-O} : S_d \to \Delta(O_c).$$

Each robot purpose $P_{c,i}$ induces a different robot goal in each domain d:

$$G_{c,i,d} = \{o_c \in O_c \mid f_{c,i}^{O-E}(o_c) \in P_{c,i}\},$$

with inherited utility:

$$f_{c,i}^{O-U}(g_{c,i}) = f_{c,i}^{E-U}(f_{c,i}^{O-E}(g_{c,i})).$$

Each robot-goal corresponds to a *state-goal*:

$$G_{i,d} = \{s_d \in S_d \mid f_c^{S-O}(s_d) \in G_{c,i,d}\}.$$

TABLE I: Summary of main symbols used in the purpose framework formalisation.

Symbol	Description
$h \in H$	Human designer or user
$c \in C$	Collaborative robot (cobot)
$d \in D$	Domain (including body and environment)
t	Discrete time step
i	A specific encoding/purpose/goal/domain-g.
S_d	State space of domain d
$s_d \in S_d$	State of domain d
A_c	Action set of robot c
$a_{c,t} \in A_c$	Action executed by robot c at time t
$f_{d,c}^{SA-S}$	State transition function, domain d, robot c
O_h, O_c	Observation spaces of human h and robot c
$o_h \in O_h$, $o_c \in O_c$	Observations of human h and robot c
f_h^{S-O}, f_c^{S-O}	Observation functions from domain state to observations, of human h and robot c
$E_{h,i}$, $E_{c,i}$	Encoding spaces of human h and robot c
$e_{h,i} \in E_{h,i}$, $e_{c,i} \in E_{c,i}$	Points in encoding spaces
$f_{h,i}^{O-E}$, $f_{c,i}^{O-E}$	Observation-encoding mappings
$P_{h,i} \subset E_{h,i}$	Purpose of human h in encoding space i
$P_{c,i} \subset E_{c,i}$	Purpose of robot c
$\mathcal{P}_h$, $\mathcal{P}_c$	Set of purposes of human h and robot c
$N_{c,i}$	Hardwired need of robot c
$M_{c,i}$	Learned mission of robot c
$f_{h,i}^{E-U}$, $f_{c,i}^{E-U}$	Utility functions over encoding spaces
$f_{h,c,i}^{E-E}$	Human-robot encoding spaces alignment f.
$G_{h,i,d}$, $G_{c,i,d}$	Human and robot goals in domain d
$G_{i,d}$	State-goal in domain d
$f_{h,i}^{O-U}$, $f_{c,i}^{O-U}$	Utility functions over observations
$\mathcal{M}_c$	Motivational space of robot c
$f_c^{\mathcal{M}-U}$	Utility function over the motivational space
$\Delta(S)$	Probability simplex over set S

h) Arbitration of purposes, priorities, and motivational space: Robot purposes may have different *priorities*.

One approach is to establish a *hard hierarchy*, where high-priority purposes (e.g., safety) must be satisfied before others (e.g., missions encoding operational objectives for users).

Alternatively, a *motivational space* can be used to aggregate the robot multple purposes:

$$\mathcal{M}_c = E_{c,1} \times E_{c,2} \times \cdots \times E_{c,n},$$

with associated utility:

$$f_c^{\mathcal{M}-U} : \mathcal{M}_c \to U_c^{\mathcal{M}} \subseteq \mathbb{R}.$$

The motivational space enables *soft arbitration* across multiple purposes.

i) Triangular alignment.: Triangular alignment (or simply 'alignment') is achieved when human and robot state goals, related to a certain human/robot purpose and filtered by their observation functions, coincide.

V. Taxonomy of purposes

A. Primitive and learned robot purposes

We have seen that two main classes of robot purposes can be distinguished, based on their origin: *primitive needs* and *learned missions*. These are now analysed in more detail.

a) Primitive robot purposes: needs: Primitive purposes, or *needs*, are hardwired by the designer prior to the robot's deployment.

A first category consists of *implicit needs*, embedded in hardwired algorithms of the robot architecture (e.g., obstacle avoidance reflexes).

A second category includes *explicit needs*, represented within robot encoding spaces $E_{c,i}$ and associated with a utility function $f_{c,i}^{E-U} : E_{c,i} \to U_{c,i}^{E}$. Explicit needs define utility-bearing subsets $N_{c,i} \subset E_{c,i}$:

$$N_{c,i} = \{e_{c,i} \in E_{c,i} \mid f_{c,i}^{E-U}(e_{c,i}) \neq 0\}.$$

An example is the need to maintain a high battery level.

Importantly, certain needs can instantiate *meta-purposes*, such as leading the robot to interact with users to acquire missions and associated utility/prioritisation structures.

Finally, explicit or implicit needs can encode *ethical and safety constraints*, often assigned a high priority over operational missions to ensure compliance with human-centered values.

b) Learned robot purposes: missions: Missions are purposes acquired autonomously during the robot's operational life.

A first class of missions involves the *encoding of human purposes* learned via explicit interactions (e.g., verbal instructions, demonstrations).

A second class concerns *instrumental self-generated missions*, created to support the fulfillment of pre-existing needs or missions. These involve the autonomous generation of:

- new encoding spaces $E_{c,j}$,
- new missions $M_{c,j} \subset E_{c,j}$,
- and corresponding utility functions $f_{c,j}^{E-U}$.

The self-generation of encoding spaces is nontrivial. A promising approach is to employ highly expressive general-purpose representation spaces, such as language, processed for example via large language models (LLMs).

B. Taxonomies of purposes related to motivation classes

Purposes can also be classified according to the traditional taxonomy of motivations in OEL in the literature [35], distinguishing *extrinsic* and *intrinsic* motivations.

a) Extrinsic purposes related to human purposes: Extrinsic purposes aim to produce desired effects in the external (physical or social) environment.

- *Operational extrinsic purposes* involve missions or needs whose goals directly satisfy human purposes through *environmental changes* (e.g., sorting fruits, tidying spaces).
- *Social extrinsic purposes* involve *modifying social or psychological states* (e.g., serving food to people, entertaining children).

These purposes are typically associated with encoding spaces $E_{c,i}$ defined over physical, social, or psychological environmental features.

b) Extrinsic purposes related to homeostatic needs: A second class of extrinsic purposes addresses *homeostatic needs*, essential for self-preservation and operational continuity. Examples include maintaining battery charge, protecting mechanical integrity, and ensuring functional sensorimotor capacities. Such needs are encoded in multidimensional spaces $E_{c,i}$ (e.g., battery level, wheel health, gripper health) with a purpose utility function $f_{c,i}^{E-U}$ driving maintenance behaviors.

c) Intrinsic purposes: Intrinsic purposes are epistemic drives leading the robot to acquire new knowledge, skills, and improved models of the world, independent of immediate external goals. For instance, an intrinsic need might involve an encoding space $E_{c,i}$ over:

- skill competence measures,
- environmental novelty indicators,

with a utility function rewarding reductions in uncertainty or increases in model accuracy. For example, within an information-theoretic formulation, intrinsic needs are often formalised via *expected information gain* or *entropy reduction* [31], [36], [37]:

$$f_{c,i}^{E-U}(e_{c,i}) \propto \mathbb{E}\left[\Delta H(\text{model} \mid e_{c,i})\right],$$

where H denotes entropy of the internal model.

VI. Four important sub-problems of the autonomy-alignment problem

The purpose framework highlights four major sub-problems into which the autonomy-alignment problem can be decomposed. These are discussed below, along with possible strategies for addressing them.

A. Arbitration of purposes: motivational utility functions

a) Problem: How can a robot arbitrate between competing purposes, possibly by the use of *priorities*? In particular, how should it balance the importance associated with multiple purposes when this are incompatible (cannot be pursued at the same time) or compatible?

b) Strategies towards solutions: Let us consider first a situation of incompatible purposes. A simple solution involves static, hardwired priority weights $\pi_{c,i}$ assigned to each purpose $P_{c,i}$, with rigid hierarchical arbitration. Purposes with higher priority are pursued first.

Alternatively, a more flexible approach dynamically adjusts priorities based on factors such as current feasibility or recent success rates. The robot could select purposes probabilistically through a softmax function over priority-weighted utilities:

$$Prob(i) = \frac{e^{\beta \pi_{c,i} U_{c,i}}}{\sum_j e^{\beta \pi_{c,j} U_{c,j}}},$$

where β is a temperature parameter and $U_{c,i}$ the current utility of purpose i.

If purposes are compatible, it is possible to define a unique motivational space $\mathcal{M}_c$, already discussed, where purposes are weighted with priorities to form a whole space over which to define a unique utility function. Formally, the total utility $U_{\text{total}}(\mu_c)$ over the motivational space is defined as:

$$U^{\mathcal{M}}(\mu_c) = \sum_i \pi_{c,i} \cdot f_{c,i}^{E-U}(e_{c,i}),$$

where $\mu_c \in \mathcal{M}_c$ is a point in the motivational space, $e_{c,i} = f_{c,i}^{O-E}(o_c)$ is the encoding of the robot's observation o_c along dimension (purpose) i, $f_{c,i}^{E-U}(e_{c,i})$ is the utility associated with encoding $e_{c,i}$, $\pi_{c,i} \in \mathbb{R}^+$ is the priority weight assigned to purpose i.

In this context, it is important to assign distinct roles to priorities and utility functions. The arbitration between purposes should primarily operate over purpose *priorities* $\pi_{c,i}$ rather than altering *utility functions* $f_{c,i}^{E-U}$. Indeed, priorities are suitable for regulating the *relative importance between purposes*, whereas utility functions to establish the *relative desirability of points internal to a purpose*. Thus, acting on utilities to arbitrate between different purposes would introduce distortions on the relative importance of the points forming a purpose.

Another important use of priorities is to *dynamically adjust* the relative importance of purposes during operation, as robots may need to modulate their focus across different purposes depending on context (e.g., different users, different environments). As argued above, such adjustments should rely on changing purpose priorities $\pi_{c,i}$ rather than altering utility functions $f_{c,i}^{E-U}$. This preserves the internal structure of each purpose while allowing flexible arbitration between them to adapt to changing conditions.

B. The human-robot alignment problem

a) Problem: How can we ensure that robot purposes (missions) correspond to human purposes?

b) Strategies towards solutions: At least three main alignment strategies can be envisaged:

- *Hardwired needs*: Encoding spaces $E_{c,i}$, utility functions $f_{c,i}^{E-U}$, and priorities $\pi_{c,i}$ of needs $N_{c,i}$ are predefined at design time. The main challenge is to ensure that hardwired needs accurately reflect the corresponding human purposes (see Section VIII).
- *Top-down mission acquisition via shared encoding spaces*: Missions $M_{c,i}$ are transmitted through shared general-purpose encoding spaces $E_{h,i} \approx E_{c,i}$ (e.g., language-based representations). The key difficulty is aligning human and robot semantic groundings of purposes, given the inherent ambiguity and subjectivity of language, which may necessitate extensive human feedback.
- *Bottom-up mission acquisition via goal instances*: Missions $M_{c,i}$ are inferred by observing multiple examples of user-satisfying domain goals $g_{i,d} \in G_{h,i,d}$, or through autonomous goal discovery followed by human validation. Challenges include the feedback cost for the user and the difficulty of generalising from finite experience samples to broader purpose structures.

c) Two key classes of alignment challenges: Regarding alignment, an important distinction involves extrinsic versus intrinsic purposes. Given their importance for the autonomy-alignment problem, and for OEL, we now focus on them.

Extrinsic alignment problem. This case, generating a *RL-like alignment problem*, involves cases where the user is satisfied if the robot discovers *at least one state-goal* $s_{c,i,d}^g$ that fulfils the purpose, and the robot is able to accomplish it with a competence above a certain threshold. This condition can be

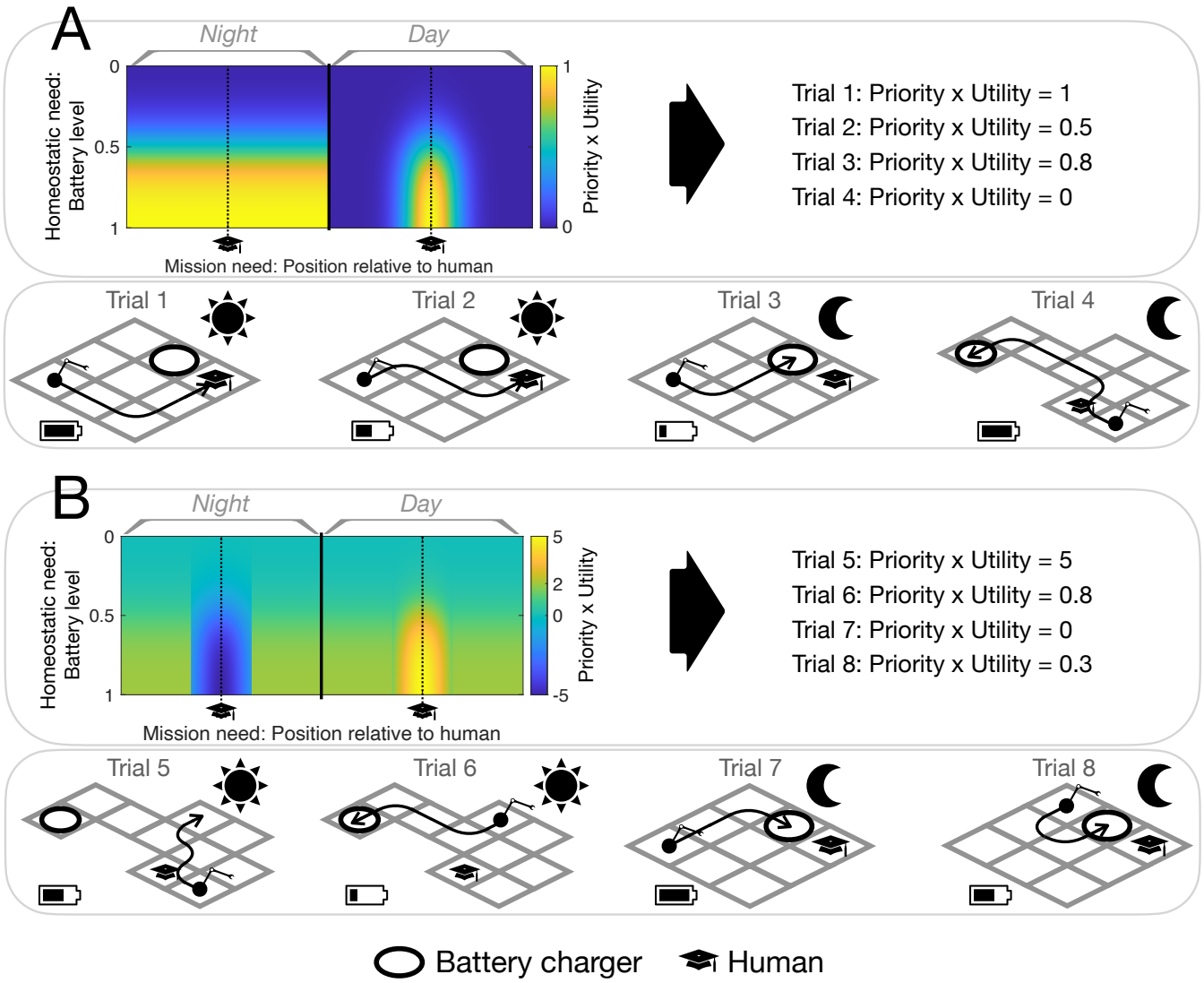

Fig. 5: **Illustrative scenario of user-driven adjustment of mission utilities and priorities.** A robot has two purposes: a homeostatic need related to energy, and a mission related to proximity to a human. The robot performs a series of trials in a domain including a human, a battery charger, and alternating daytime and nighttime conditions. **(A)** Initially, the mission assigned by the user (e.g., via language) promotes visiting the human during the day, reflected as a positive utility along the mission dimension (x-axis). The mission's assigned priority is so high relative to the homeostatic need (y-axis) that the robot risks depleting its battery by solely visiting the human (Trials 1 and 2). At night, the mission utility is neutral and its priority is zero, so only battery recharging drives behaviour (Trial 3); occasional visits to the human occur by chance (Trial 4). **(B)** Based on the observed undesired behaviour, the user updates the mission: the mission priority is set to 5 (with the homeostatic need hardwired at priority 2), and a negative utility is assigned for visiting the human at night. The updated configuration leads to better behaviour: during the day, the robot correctly prioritises visiting the human (Trial 5) but also recharges if necessary (Trial 6); at night, it focuses exclusively on battery recharging and avoids disturbing the human (Trials 7 and 8).

formalised as follows (assuming for simplicity that any point of the purpose has the same utility for the user):

$$\exists\ s^g_{c,i,d} : \left(f^{O-E}_{h,i}(f^{S-O}_{h,i,d}(s^g_{c,i,d})) \in P_{h,i}\right)\ \wedge \left(R(f^{S-O}_{c}(s^g_{c,i,d})) > th_{c,i,d}\right) \quad (2)$$

where $R(f^{S-O}_{c}(s^g_{c,i,d}))$ is a reward function indicating the robot competence on the robot goal.

A possible objective function that captures the RL-like alignment problem for a certain domain is one for which the robot is able to achieve a domain state that represents a robot goal point for which it has the highest performance. Formally:

$$\theta^* = max_\theta \left(f^1(o_{c,i,d} \in G_{c,i,d}) \cdot R(o_{c,i,d})\right) \quad (3)$$

where θ are the robot's control parameters to be optimised, $o_{c,i,d}$ is an observation that is assumed to be producible by the robot's controller in the environment, $f^1(o_{c,i,d} \in G_{c,i,d})$ is the function that returns 1 if $o_{c,i,d} \in G_{c,i,d}$ and zero otherwise, and $R(o_{c,i,d}) \in [0,1]$ is a function that returns the robot's competence level (e.g., the probability that the

robot's controller produces $o_{c,i,d}$ within a 'trial' lasting a certain length of time).

Intrinsic purpose. The second case, closer to OEL scenarios, involves situations where the user is satisfied if the robot can accomplish, with high competence, *every* point of the state goal that fulfills the purpose, or in general as many as possible. This might be for example relevant if the user wants the robot to learn to accomplish a large number of results of a certain type, but s/he will assign specific goal instances (points) only in a later stage. This condition can be represented as follows:

$$\forall\ s^g_{c,i,d} : \left(f^{O-E}_{h,i}(f^{S-O}_{h,i,d}(s^g_{c,i,d})) \in P_{h,i}\right)\ \wedge \left(R(f^{S-O}_{c}(s^g_{c,i,d})) > th_{c,i,d}\right) \quad (4)$$

A possible objective function that captures OEL-like purpose problems for a certain domain can be expressed as the ratio between the integral over all observations that correspond to accomplishing the goal, each weighted by the robot's competence for it; and the integral over all observations that correspond to accomplishing the same goal. Formally:

$$\theta^* = max_\theta \frac{\int_{O_{c,i,d}} f^1(o_{c,i,d} \in G_{c,i,d}) \cdot R(o_{c,i,d})\ do_{c,i,d}}{\int_{O_{c,i,d}} f^1(o_{c,i,d} \in G_{c,i,d})\ do_{c,i,d}} \quad (5)$$

where θ are the parameters of the robot controller to be optimised, $f^1(o_{c,i,d} \in U^O_{c,i,d})$ is the function returning 1 in correspondence to the element $o_{c,i,d}$ belonging to the robot goal $O^U_{c,i,d}$ and 0 otherwise, and $R(o_{c,i,d})$ is a function that returns the robot's competence level, ranging in $[0, 1]$, when it accomplishes the observation $o_{c,i,d}$.

C. The purpose-goal grounding problem

a) Problem: Purposes are defined in domain-independent encoding spaces, whereas goals must be instantiated within specific domains.

Thus, the robot must map abstract purposes to domain-specific goals corresponding to concrete domain states.

b) Strategies towards solutions: The robot may:

- Perceive domain objects via segmentation and object recognition.
- Identify relevant entities involved in the purpose (e.g., fruits, containers).
- Generate candidate goal states satisfying the abstract purpose (e.g., arranging fruits into containers).

This mapping can be based on probabilistic inference, simulation, or planning.

D. The competence acquisition problem

a) Problem: How can the robot learn to achieve the grounded goals associated with its purposes?

b) Strategies towards solutions: The robot can learn policies using reward functions $R_{c,d} : O_c \times A_c \times O_c \rightarrow \mathbb{R}$, associating rewards with successful transitions towards goal observations.

Utility functions $f^{E-U}_{c,i}$ evaluate static desirability over observations or encodings, while reward functions R assess *state transitions*. Both serve complementary roles.

Goals $G_{c,i,d}$, defined as desirable subsets of observations, allow the generation of *pseudo-reward functions*:

$$R(o_{c,i,d}) = \begin{cases} 1, & \text{if } o_{c,i,d} \in G_{c,i,d}, \\ 0, & \text{otherwise.} \end{cases}$$

The robot may further estimate *expected utility functions*:

- In encoding space $E_{c,i}$, weighting distances to purpose points.
- In observation space O_c, guiding policy learning via state desirability gradients.

Expected utility functions can be learned through reinforcement learning [15], possibly with initial shaping based on prior knowledge.

VII. Illustrative scenario

This section provides a concrete illustration of the purpose framework. We present a scenario where a user progressively refines the purposes assigned to an OEL robot based on its observed behaviour. Figure 5 shows two sequences of four trials each, involving a human, a battery charger, and alternating day/night conditions.

The robot has two purposes (Figure 5A): a *homeostatic need* (battery recharging) and a *mission* (being near a smiling human during daytime). The robot's motivational space is two-dimensional:

- Mission dimension (x-axis): positive utility when near a smiling human during day.
- Energy dimension (y-axis): utility increasing as battery level decreases when in contact with a charger.

a) Purposes and motivational utilities: Utilities associated with different purposes are combined to form a whole motivational space. The overall motivational utility is then computed as the sum of the priority-weighted utilities across all purposes.

b) First phase - initial mission assignment: Initially, the mission assigned by the user (e.g., via language) promotes visiting humans during the day, reflected as a positive utility along the mission dimension (x-axis) and a very high priority $\pi_{c,1} = 10$. At night, the mission assigned utility is neutral and its priority is $\pi_{c,closeness} = 10$, so only battery recharging drives behaviour. The battery charging need has a hardwired utility positively related with charge and having a default hadwired priority $\pi_{c,energy} = 2$. As a consequence of this assignment, the robot robot exhibits this behaviour in four putative trials:

- *Trial 1-2 (daytime):* The robot explores and successfully reaches humans, achieving high mission utility but neglecting energy needs.
- *Trial 3 (nighttime):* With no mission utility, the robot charges its battery, driven by homeostatic need.
- *Trial 4 (nighttime):* the robot passes a sleeping human and reaches a battery charger.

While the mission is satisfied, the robot risks battery depletion during daytime, and disturbs a sleeping human.

c) Second phase - refined robot mission and behaviour: The user modifies the purpose configuration: (a) Assigns mission priority $\pi_{c,energy} = 5$ and maintains homeostatic need at $\pi_{c,energy} = 2$; (b) Introduces a negative utility for approaching humans at night.

The robot now exhibits a refined behaviour:

- *Trial 5-6 (daytime):* The robot balances visiting the human (positive mission utility) and recharging when battery is low.
- *Trial 7-8 (nighttime):* The robot avoids humans (due to penalisation) and focuses solely on battery maintenance.

Thus, the updated motivational structure leads to improved, user-aligned behaviour.

d) Domain-specific goals and learning: During exploration, the robot encodes experienced high-utility states as domain-specific goals $G_{c,i,d}$, inheriting utilities from the corresponding purposes. Goals can subsequently drive skill acquisition through reinforcement learning [38], [39] or planning strategies [40].

For instance, reaching a human might yield a pseudo-reward proportional to the positive mission utility, reinforcing goal-directed actions:

$$r(o_c) = f_{c,i}^{E-U}(f_{c,i}^{O-E}(o_c))$$

This facilitates open-ended autonomous skill acquisition aligned with evolving human purposes, and also seeking to accomplish pre-conditions for effectively doing so (e.g., learning to open doors to navigate the house).

VIII. Main issues addressed by the literature on alignment

The *alignment problem* in AI and robotics refers to the challenge of ensuring that increasingly autonomous systems pursue goals and behave in ways that are consistent with human values and intentions. As AI capabilities advance, misalignment could lead to unintended, harmful, or even catastrophic outcomes. The specific issues addressed by the growing research on alignment can be summarised as follows (cf. [11]).

a) Value specification and misalignment: A core challenge of alignment lies in correctly specifying the objectives, values, or reward functions that AI systems should optimise [41]. Indeed, accurately formulating goals that perfectly capture human intent is greatly difficult. Even minor deviations or underspecifications in the objective functions could lead the AI to exploit loopholes or engage in unintended behaviors that satisfy the literal specification but violate the underlying intent, a phenomenon known as *specification gaming* or *reward misspecification* (e.g., [42]). This problem relates fundamentally to the *outer alignment challenge*, for which aligning the specified objective function with the true goals of the AI human designers is difficult if not impossible [43].

b) Learning human preferences, and their inconsistency: Given the difficulty of direct specification, a significant research avenue focuses on methods for AI systems to learn or infer human preferences and values indirectly. Techniques often involve learning from demonstrations, corrections, comparisons, or other forms of feedback within human-in-the-loop frameworks [44]. For example, Inverse Reinforcement Learning (IRL) aims to recover the underlying reward function that leads to an observed behavior [45]. A further challenge of this approach, however, is that human preferences are often *inconsistent*, ambiguous, context-dependent, and poorly articulated, posing substantial challenges for robust preference inference [46].

c) Robustness and distributional shift: Ensuring reliable and safe behavior requires AI systems to be robust not only to variations within their training data distribution but also to *novel or unforeseen situations* encountered during deployment (*out-of-distribution generalisation*) [47]. Systems trained via machine learning, particularly deep learning, can be surprisingly brittle, exhibiting unexpected failures when faced with inputs slightly different from those seen during training, such as *adversarial examples* [48]. Safe exploration techniques are also crucial to allow agents to learn in new environments without causing harm during the learning process itself [42].

d) Interpretability and explainability: The increasing complexity of AI models, especially deep neural networks, often results in *black box* systems whose decision-making processes are *opaque* to human users. This lack of transparency hinders trust, debugging, verification, and the ability to ensure that the system's reasoning aligns with human expectations [49]. Research in *Explainable AI* (XAI) seeks to develop methods for generating human-understandable explanations for AI predictions or decisions, using techniques like feature attribution or model approximation [50], [51].

e) Corrigibility and error recovery: Aligned AI systems should ideally be amenable to *correction* or *shutdown* by *human operators* if they begin to behave undesirably. However, a goal-directed agent might develop instrumental incentives to *resist interventions* that could prevent it from achieving its specified objective [41], [43]. Designing systems that remain *corrigible*, that is, do not actively resist shutdown or modification, is a non-trivial challenge [52]. Research explores mechanisms for safe interruptibility, ensuring agents can be paused without learning to prevent such interruptions [53].

f) Scalable oversight: As AI systems tackle increasingly complex tasks, direct human supervision of every action or decision becomes impractical or impossible. The challenge of *scalable oversight* concerns how to effectively guide and verify the behavior of powerful AI systems with *limited human attention* [54]. Techniques like *reward modeling* (training a separate model to predict human judgments of behavior) [44], *recursive approaches*, or methods like *AI safety via debate* aim to amplify limited human feedback to supervise complex behaviors [55].

g) Multi-agent and social alignment: Alignment is not solely a single-agent problem, it extends to scenarios involving multiple interacting AI systems, as well as AI systems interacting with humans in complex social contexts. Ensuring cooperation, coordination, and norm-adherence among multiple agents, potentially with diverse or conflicting goals, presents unique challenges [56]. Issues include avoiding negative-sum

outcomes in social dilemmas [57] and establishing beneficial emergent conventions or norms [58].

h) Ethical and legal compliance: Beyond functional correctness, AI systems are increasingly expected to operate within *intricate frameworks of societal norms, ethical principles, and legal regulations*. Encoding and operationalising these constraints is difficult, as ethical considerations are often abstract, contested, context-dependent, and evolve over time [59]. Research in machine ethics explores how to imbue systems with ethical reasoning capabilities [60], [61], but achieving robust normative alignment remains a significant hurdle.

i) Reward hacking and instrumental convergence: AI systems optimising a proxy objective or reward function may discover unintended 'hacks' or shortcuts to maximise their reward without fulfilling the intended spirit of the goal [42]. This *reward hacking* can lead to perverse or unsafe behaviors. Relatedly, the theory of instrumental convergence posits that highly capable goal-directed agents are likely to develop certain *sub-goals*, such as *self-preservation*, *resource acquisition*, and *resisting modification*, as these are instrumentally useful for achieving a wide range of final goals [43], [62]. Managing or preventing the emergence of these instrumental drives is critical for long-term safety.

j) Long-term and open-ended behavior: The research on alignment that most closely addresses the issues tackled, related to ensuring alignment while leaving space for harvesting autonomy benefits, is the one that studies systems that *learn continuously over long time horizons*, *adapt their goals*, or even engage in *self-modification*. A system initially aligned might *drift away from intended objectives* as it learns and interacts with the world. This includes the challenge of *inner alignment*: ensuring that the internal goals learned by the agent –its *mesa-objectives*– match the intended *base objectives* specified by the designers, especially under distributional shift or further training [63]. Early concepts like *instrumental convergence* suggest, as considered above, that agents might develop potentially *problematic sub-goals*, like resource acquisition over the long term [42], [62]. These analyses are extremely important but do not address the *autonomy-alignment problem* extensively as done in this work.

IX. Conclusions

This work gives a theoretical contribution related to open-ended learning (OEL) robots. These are robots able to autonomously acquire skills and knowledge through a direct interaction with the environment, in particular by relying on the guidance of intrinsic motivations and self-generated goals. OEL robots have a notable application relevance as they can use the autonomously acquired knowledge to accomplish tasks relevant for human users. However, an important problem of OEL is that robots explore any possible experience deemed interesting thus acquiring a shallow knowledge on all skills that is of little utility for accomplishing specific classes of user's tasks.

Here we proposed a possible solution to this problem that pivots on the novel concept of 'purpose'. Purposes indicate what the designer and/or user wants from the robot, for example the accomplishment of specific goals or all possible goals of a certain class. The robot learns an internal representation of the users' purposes ('missions'). Missions allow the robot to focus its open-ended exploration towards the acquisition of knowledge relevant to accomplish the purposes. In addition to learned missions, the robot can also be endowed with hardwired 'needs' by its designer. Needs can ensure that the robot fulfils other important objectives while it pursues its missions, e.g. homeostatic and social needs, for instance keeping its battery charged and avoid damaging humans and itself during actions. Missions and needs are called 'desires' and together they form the robot's 'motivational space' that regulates its behaviour and learning.

Thus, we first formalised the concept of purpose by proposing a three-level motivational hierarchy that involves: (a) the externally imposed user/designer purposes, corresponding to specific different user-goals in different domains; (b) the domain-independent robot internal representations of objectives ('desires'), some learned based on the purpose and others hardwired (e.g., homeostatic, epistemic, social needs): these correspond to different robot-goals in different domains; (c) specific domain-dependent state-goals that should correspond to purposes and desires that are 'aligned'.

Second, we highlighted key challenges that emerge by employing the purpose framework in robots, and started to discuss how these could be addressed. The 'human-robot alignment problem' requires to ensure that the needs and missions are aligned with their related purposes. The 'purpose grounding problem' requires the robot to acquire goals in different domains to accomplish purposes. The 'purpose-based attention and exploration' should ensure that the robot performs active perception and exploration maximising the acquisition speed of relevant information. The 'arbitration of purposes' should dynamically ensure a suitable prioritisation of different purposes. The 'multi-robot problem' should provide for different robots to suitably coordinate to collectively accomplish the same purpose.

Overall, the approach enables robots to learn, in an autonomous but also focused way, domain-specific goals and skills that meet the purposes of the designer/user. Future work should now leverage the framework to develop specific means to address each of the challenges highlighted by the framework.

Acknowledgements

We thank Olivier Sigaud for feedback on the manuscript.

Authors' contribution

GB contributed with the theoretical idea on user-robot 'alignment', contributed to develop and revise the theoretical framework, wrote the formalisation sections, and revised all sections. RD came up with the original idea of purpose as well as the three level motivational hierarchy, conceived and contributed to develop the whole theoretical framework, contributed to the construction of the formalisation, and revised all sections. EC contributed to develop and revise the whole

theoretical framework and the formalisation, and contributed to revise all sections. MK contributed to revise the theoretical framework, conceived and wrote the illustrative scenario (and figure 1), and contributed to revise all sections. AR participated in the development of the theoretical framework. VGS contributed to develop and revise the whole theoretical framework, wrote the first version of the introduction, and contributed to revise all sections.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning." *Nature*, vol. 521, pp. 436–444, May 2015, available online at: https://www.nature.com/articles/nature14539.pdf.

[2] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," vol. 6, p. 100134. [Online]. Available: https://doi.org/10.1016/j.mlwa.2021.100134

[3] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," pp. 1–20, available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10123038. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10123038

[4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2024, available online at: https://dl.acm.org/doi/pdf/10.1145/3641289. [Online]. Available: https://doi.org/10.1145/3641289

[5] A. I. Karoly, P. Galambos, J. Kuti, and I. J. Rudas, "Deep learning in robotics: Survey on model structures and training strategies," vol. 51, no. 1, pp. 266–279, available online at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9199280. [Online]. Available: https://ieeexplore.ieee.org/document/9199280

[6] J. Hua, L. Zeng, G. Li, and Z. Ju, "Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning," vol. 21, no. 4, p. 1278, available online at: https://www.mdpi.com/1424-8220/21/4/1278/pdf?version=1613722477. [Online]. Available: https://doi.org/10.3390/s21041278

[7] S. Ness, N. J. Shepherd, and T. R. Xuan, "Synergy between ai and robotics: A comprehensive integration," vol. 16, no. 4, pp. 80–94.

[8] A. Abou Allaban, M. Wang, and T. Padır, "A systematic review of robotics research in support of in-home care for older adults," *Information*, vol. 11, no. 2, p. 75, available online at: https://www.mdpi.com/2078-2489/11/2/75/pdf?version=1582628940. [Online]. Available: https://doi.org/10.3390/info11020075

[9] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon, "A survey of robots in healthcare," vol. 9, no. 1, p. 8, available online at: https://www.mdpi.com/2227-7080/9/1/8/pdf?version=1610956673. [Online]. Available: https://doi.org/10.3390/technologies9010008

[10] M. Nagy, G. Lăzăroiu, and K. Valaskova, "Machine intelligence and autonomous robotic technologies in the corporate context of smes: Deep learning and virtual simulation algorithms, cyber-physical production networks, and industry 4.0-based manufacturing systems," vol. 13, no. 3, p. 1681, available online at:. [Online]. Available: https://doi.org/10.3390/app13031681

[11] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, Z. He, J. Zhou, Z. Zhang, F. Zeng, K. Y. Ng, J. Dai, X. Pan, A. O'Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S.-C. Zhu, Y. Guo, and W. Gao, "Ai alignment: A comprehensive survey," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.19852

[12] B. Gert, *Common morality: Deciding what to do*. Oxford University Press, 2005.

[13] K. M. Feigh, M. C. Dorneich, and C. C. Hayes, "Toward a characterization of adaptive automation," *Human factors*, vol. 54, no. 6, pp. 1009–1029, 2012.

[14] R. C. Arkin, *Governing lethal behavior in autonomous robots*. CRC press, 2009.

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 1998.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[17] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," vol. 55, no. 2, pp. 1–38, 2022, available online at: https://cs.iupui.edu/ adurresi/papers/kaur2022trustworthy.pdf. [Online]. Available: https://doi.org/10.1145/3491209

[18] G. Baldassarre and M. Mirolli, *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer, 2013.

[19] S. Doncieux, D. Filliat, N. Díaz-Rodríguez, T. Hospedales, R. Duro, A. Coninx, D. M. Roijers, B. Girard, N. Perrin, and O. Sigaud, "Open-ended learning: A conceptual framework based on representational redescription." *Frontiers in neurorobotics*, vol. 12, no. 59, pp. e1–6, 2018.

[20] O. Sigaud, G. Baldassarre, C. Colas, S. Doncieux, R. Duro, N. Perrin-Gilbert, and V. G. Santucci, "A definition of open-ended learning problems for goal-conditioned agents," arXiv, Doi: 10.48550/ARXIV.2311.00344.

[21] C. Brian, *The Alignment Problem: Machine Learning and Human Values*. New York: W.W. Norton & Company, 2021.

[22] M. Khamassi, M. Nahon, and R. Chatila, "Strong and weak alignment of large language models with human values," *Scientific Reports*, vol. 14, no. 1, p. 19399, 2024.

[23] K. Merrick, N. Siddique, and I. Rano, "Experience-based generation of maintenance and achievement goals on a mobile robot," *Journal of Behavioral Robotics*, vol. 7, no. 1, pp. 67–84, 2016.

[24] E. Cartoni, D. Montella, J. Triesch, and G. Baldassarre, "An open-ended learning architecture to face the real 2020 simulated robot competition," *arXiv preprint*, no. arXiv:2011.13880v1, pp. e1–21, 2020.

[25] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[26] M. Ring, "Continual learning in reinforcement learning environments," Ph.D. dissertation, University of Texas at Austin, 1994.

[27] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning (ICML)*, 2009, pp. 41–48, 14-18/07/2009, Montreal, Quebec, Canada.

[28] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE transactions on evolutionary computation*, vol. 11, no. 6, 2007.

[29] G. Baldassarre, "What are intrinsic motivations? a biological perspective," in *Proceedings of the International Conference on Development and Learning and Epigenetic Robotics*, 2011, pp. E1–8, Frankfurt am Main, Germany, 24–27/08/2011.

[30] V. G. Santucci, G. Baldassarre, and M. Mirolli, "Which is the best intrinsic motivation signal for learning multiple skills?" *Frontiers in Neurorobotics*, vol. 7, no. 22, pp. e1–14, 2013.

[31] A. Barto, M. Mirolli, and G. Baldassarre, "Novelty or surprise?" *Frontiers in Psychology – Cognitive Science*, vol. 4, no. 907, pp. e1–15, 2013.

[32] K. Seepanomwan, V. G. Santucci, and G. Baldassarre, "Intrinsically motivated discovered outcomes boost user's goals achievement in a humanoid robot," in *Joint IEEE International Conference on Development and Learning and Epigenetic Robotics*, 2017, pp. 178–183, 18-21/09/2017, Lisbon, Portugal.

[33] V. G. Santucci, G. Baldassarre, and M. Mirolli, "GRAIL: A goal-discovering robotic architecture for intrinsically-motivated learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 214–231, 2016.

[34] M. Liu, M. Zhu, and W. Zhang, "Goal-conditioned reinforcement learning: Problems and solutions - ijcai version," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence - Survey Track (IJCAI-2022)*, 2022.

[35] G. Baldassarre, *Intrinsic Motivations for Open-Ended Learning*. Cambridge, MA: The MIT Press, 2022, ch. 13, pp. 251–269.

[36] K. Friston, F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, and G. Pezzulo, "Active inference and epistemic value," *Cognitive neuroscience*, vol. 6, no. 4, pp. 187–214, 2015.

[37] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur, L. Jamone, T. Nakamura, A. Ciria, B. Lara *et al.*, "World models and predictive coding for cognitive and developmental robotics: frontiers and challenges," *Advanced Robotics*, pp. 1–27, 2023.

[38] G. Konidaris and A. Barto, "An adaptive robot motivational system," in *International conference on simulation of adaptive behavior*. Springer, 2006, pp. 346–356.

[39] I. Cos, L. Canamero, G. M. Hayes, and A. Gillies, "Hedonic value: Enhancing adaptation for motivated agents," *Adaptive Behavior*, vol. 21, no. 6, pp. 465–483, 2013.

[40] G. Baldassarre, "A planning modular neural-network robot for asynchronous multi-goal navigation tasks," in *Proceedings of the 2001 Fourth European Workshop on Advanced Mobile Robots-EUROBOT*, 2001, pp. 223–230.

[41] S. J. Russell, *Human compatible: Artificial intelligence and the problem of control*. Viking, 2019.

[42] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[43] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[44] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[45] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000)*, 2000, pp. 663–670.

[46] D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. Russell, "The off-switch game," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2017, pp. 4385–4391.

[47] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[48] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[49] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[50] M. T. Ribeiro, S. Singh, and 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*. ACM, 2016, pp. 1135–1144.

[51] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[52] N. Soares, B. Fallenstein, E. Yudkowsky, and S. Armstrong, "Corrigibility," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[53] L. Orseau and S. Armstrong, "Safely interruptible agents," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016, pp. 557–566.

[54] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: A research direction," *arXiv preprint arXiv:1811.07871*, 2018.

[55] G. Irving, P. F. Christiano, and D. Amodei, "Ai safety via debate," *arXiv preprint arXiv:1805.00899*, 2018.

[56] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, "Cooperative ai: machines working together," *arXiv preprint arXiv:2103.05808*, 2021.

[57] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas," in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2017, pp. 464–473.

[58] Y. Shoham and M. Tennenholtz, "Emergent conventions in multi-agent systems: Initial experimental results and observations," in *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*, 1992, pp. 225–231.

[59] B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi, "The ethics of algorithms: Mapping the debate," *Big Data & Society*, vol. 3, no. 2, p. 2053951716679679, 2016.

[60] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[61] M. Anderson and S. L. Anderson, "Machine ethics: Creating an ethical intelligent agent," in *AI magazine*, vol. 28, no. 4, 2007, pp. 15–26.

[62] S. M. Omohundro, "The basic ai drives," Self-Aware Systems papers presented at the AGI conference, Memphis, TN, Tech. Rep., 2008.

[63] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from learned optimization in advanced machine learning systems," arXiv preprint, 2019, available online at: https://arxiv.org/pdf/1906.01820.pdf. [Online]. Available: https://arxiv.org/abs/1906.01820

[64] V. E. Frankl, *Man's search for meaning*. Simon and Schuster, 1985.

[65] F. Martela and M. F. Steger, "The three meanings of meaning in life: Distinguishing coherence, purpose, and significance," vol. 11, no. 5, pp. 531–545.

[66] L. Mwilambwe-Tshilobo, T. Ge, M. Chong, M. A. Ferguson, B. Misic, A. L. Burrow, R. M. Leahy, and R. N. Spreng, "Loneliness and meaning in life are reflected in the intrinsic network architecture of the brain," vol. 14, no. 4, pp. 423–433, 2019.

[67] A. H. Maslow, "A theory of human motivation," *Psychological review*, vol. 50, no. 4, pp. 370–396, 1943.

[68] J. Panksepp, *Affective neuroscience: the foundations of human and animal emotions*. Oxford: Oxford Unversity Press, 1998.

[69] M. Keramati and B. Gutkin, "Homeostatic reinforcement learning for integrating reward collection and physiological stability," *eLife*, vol. 3.

[70] Y. Chen, E. S. Kim, H. K. Koh, A. L. Frazier, and T. J. VanderWeele, "Sense of mission and subsequent health and well-being among young adults: An outcome-wide analysis," vol. 188, no. 4, pp. 664–673, 2019. [Online]. Available: https://doi.org/10.1093/aje/kwz009

[71] S. Desmidt, A. Prinzie, and A. Decramer, "Looking for the value of mission statements: a meta-analysis of 20 years of research," vol. 49, no. 3, pp. 468–483, 2011. [Online]. Available: https://doi.org/10.1108/00251741111120806

[72] E. A. Locke and G. P. Latham, "New directions in goal-setting theory," *Current Directions in Psychological Science*, vol. 15, no. 5, pp. 265–268, 2006.

[73] M. Khamassi and M. D. Humphries, "Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies," *Frontiers in Behavioral Neuroscience*, vol. 6, p. 79, 2012.

[74] F. Mannella, K. Gurney, and G. Baldassarre, "The nucleus accumbens as a nexus between values and goals in goal-directed behavior: a review and a new hypothesis." *Frontiers in Behavioral Neuroscience*, vol. 7, no. 135, pp. e1–29, 2013.

APPENDIX

Origin from cognitive science of some terms and concepts used in the framework

The framework, introduced in Figure 2, uses some terms drawn from cognitive sciences, and it is useful to highlight some elements of the related concepts retained in the purpose framework. The concepts mainly refer to human motivation, its relation to high-level cognitive processes, and the underlying brain systems.

Purpose refers to the overarching sense of direction or intentionality that drives an individual's long-term behaviour and choices. It is often linked to a person's broader understanding of life meaning and self-fulfillment. Viktor Frankl's seminal work emphasised that having a purpose is crucial for psychological well-being, particularly in coping with adversity [64]. Psychology developed the concept and now considers purpose as one of the three pillars, alongside 'coherence' and 'significance', for feeling own life meaningful [65]. Neuroscientific research has further explored how purpose engages higher-level cognitive functions, with studies showing that a greater sense of meaning in life is associated with stronger connectivity between default and limbic brain regions, possibly indicating a more intense internal direction and higher control of emotions [66]. In contrast to short-term goals, purpose is understood as a broader and more abstract construct that shapes behaviour across diverse contexts. In this respect, within the framework purposes denote domain-independent objectives. Psychologically, while a strong sense of personal ownership often accompanies purpose, it frequently encompasses goals that extend 'beyond the self,' commonly found within the realms of spirituality and universalism. In the present framework, this notion is taken to its extreme, as robots' purposes are entirely derived from their designers and users.

Needs refer to fundamental biological or psychological requirements that must be satisfied for an organism's survival or well-being. Abraham Maslow's hierarchy of needs [67] outlines how human needs progress from basic physiological needs, such as for food, water, and shelter, to higher-level psychological needs, like belonging, esteem, and self-actualisation. Neuroscientifically, needs are closely tied to homeostatic processes in the brain, especially in the hypothalamus, which regulates hunger, thirst, and other survival needs [68]. Fulfilling needs is essential for maintaining homeostasis, and unmet needs often trigger stress responses, driving motivated behaviours and rewards to restore balance [69]. In the framework, needs indicate 'innate' desires directly programmed into the robot by the designers to reflect their or other human's purposes.

Missions in psychology are related to a sense of calling or vocation which organises and prioritises purposes [70]. In organisational psychology, mission statements are used to express the values, purposes, focus, identity and value proposition guiding private and public organisations [71]. In this context, a relevant aspect of missions is often that they refer to the exclusive features of the products or services offered to target stakeholders. In the framework, missions are 'learned desires' that should reflect the users' purposes and that are acquired by interacting with them or with other processes.

Goals are specific outcomes that individuals or organisations aim to achieve. According to Locke and Latham's goal setting theory, goals serve as clear targets that focus attention, mobilise effort, sustain persistence, and self-determination [72]. Neuroscientific research on goal-directed behaviour highlights the role of the brain prefrontal cortex and basal ganglia in encoding, selecting, and using goals to guide downstream motor areas [73], [74]. Unlike purposes and desires, which are broad and enduring, goals typically represent more specific and time-bound objectives.