
Humans Coexist, So Must Embodied Artificial Agents

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 This position paper introduces the concept of coexistence for embodied artificial
2 agents and argues that it is a prerequisite for long-term, in-the-wild interaction
3 with humans. Contemporary embodied artificial agents excel in static, predefined
4 tasks but fall short in dynamic and long-term interactions with humans. On the
5 other hand, humans can adapt and evolve continuously, exploiting the situated
6 knowledge embedded in their environment and other agents, thus contributing
7 to meaningful interactions. We take an interdisciplinary approach at different
8 levels of organization, drawing from biology and design theory, to understand
9 how human and non-human organisms foster entities that coexist within their
10 specific environments. Finally, we propose key research directions for the artificial
11 intelligence community to develop coexisting embodied agents, focusing on the
12 principles, hardware and learning methods responsible for shaping them.

13 1 Introduction

14 Contemporary artificial intelligence (AI) systems have shown remarkable performance across diverse
15 tasks such as the high-quality generation of data (image, text, video) [54, 2, 92], the creation of
16 interactive world models [21, 4], and outperforming humans in complex decision-making tasks [127,
17 143, 140]. Fundamentally, three ingredients have been mostly responsible for this recent surge in
18 performance: the creation of large-scale models [141, 32], the curation (or creation) of internet-scale
19 datasets [123, 53] and a computationally-intensive offline training process [114, 157, 20]. This recipe
20 has also been replicated for real-world robotic systems, resulting in the creation of large-scale datasets
21 of expert-level interaction data in the real-world [106] and in simulation environments [145]. This
22 approach has led to progresses in learning generalist robotic policies, able to perform a wide variety
23 of manipulation and navigation tasks [14, 156].

24 As a community, we now envision concrete use cases of *embodied artificial agents*¹ for human
25 interaction². Despite their remarkable progress in controlled environments [18], embodied agents
26 still struggle to gain a foothold in-the-wild scenarios [6]. Rodney Brooks’ famous quip, “The world
27 is its own best model” [110] is often used to encapsulate the problem of conceiving and deploying
28 embodied artificial agents in the real-world [12]. However, we highlight that this challenge does
29 not only emerge from the complex and dynamic nature of the real-world; it also stems from the
30 constant tendency of viewing the real-world as an optimization problem [128]. Interaction *in-the-wild*,
31 instead, is co-constructed with the humans *in-the-wild* [42], which is at odds with the dominant
32 *problematize-solve-optimize-deploy* workflow of the contemporary AI community [59].

¹We follow Paolo et al. [107] that defines embodied artificial agents as “agents that interact with their physical environment, emphasizing sensorimotor coupling and situated intelligence”. Throughout this paper, we use the terms *agent*, *embodied agent*, and *embodied artificial agent* interchangeably for simplicity.

²In Appendix A we discuss the scope of interaction in the context of embodied agents.

33 **We argue that the current approach to agent design is unsuitable for long-term, in-the-wild**
 34 **interaction with humans.** In Section 2, we discuss why current embodied artificial agents are unable
 35 to cope with the strong dynamic nature of human interaction and the agents’ inability to participate
 36 in its ongoing evolution. **We emphasize the need for a new paradigm for coexisting embodied**
 37 **agents:** mutable systems capable of continuously leveraging situated knowledge of both the user and
 38 the environment, highlighted in Figure 1, to establish meaningful and reciprocal interactions with the
 39 elements of its system. In Section 3, we formally define coexistence and its properties (situatedness
 40 and mutability) in the context of embodied artificial agents.

41 Our formal definition is complemented by a more practical approach for such agents to coexist in-
 42 the-wild. **In Section 4, we take an interdisciplinary perspective to understand how human and**
 43 **non-human organisms foster entities that coexist within their specific environments.** Therefore,
 44 we look to evolutionary biology and design theory, two fields that are epistemically grounded in the
 45 real-world³. There we highlight how biological organisms leverage properties of the real-world to
 46 take form during development (converge), and evolve in times of environmental changes (diverge).
 47 Similarly, we highlight how the double diamond design process (depicted in Figure 2a) has been an
 48 indispensable tool that has allowed designers and engineers to physically explore (diverge) and refine
 49 (converge) creative solutions. This section is not a call for the underlying systems of artificial agents
 50 to more closely resemble those of biological ones (for example, as in Darlow et al. [30]). Rather, to
 51 acknowledge the dynamic and interconnected nature of biological systems, as well as the existing
 52 methodologies in design theory that are adept with handling similar complexities. Taken together,
 53 they provide direction on building autonomous agents and our relationships with them.

54 **In Section 5, we highlight six key research directions for the AI community to develop coexisting**
 55 **agents.** We focus on the learning methods that enable coexistence, the hardware that sustains it,
 56 and the principles responsible for shaping it. Additionally, we discuss the ethical considerations in
 57 designing embodied agents that coexist with humans and play a role in shaping the future of human
 58 interactions. Finally, in Section 6 we contrast coexistence with alternative (predominant) viewpoints
 59 on embodied artificial agents. We see our work as a bridge, enabling the AI community to actively
 60 engage with the design research community in forging a path toward coexisting embodied agents.

61 2 Current Embodied Agents Exist

62 Recent advancements in perception, learning and hardware systems have enabled embodied agents
 63 to successfully perform complex actions in unstructured environments [54, 2, 92]. We praise
 64 these advancements and believe that the current paradigm, based on multimodal foundation models
 65 for perception, reasoning and interaction, will be sufficient for these agents to *exist* with humans
 66 and within their environments. However, we argue that the disregard of the issues pertaining to
 67 current embodied agents can have technical and cultural repercussions when employed widely in our
 68 societies. In particular, we focus on two fundamental properties of these agents: their *stagnant* nature,
 69 a consequence of having their abilities fixed at a specific moment in time, and their *generic* nature,
 70 due to their instantiation based solely on large amounts of pre-collected data. As current embodied
 71 agents are stagnant and generic, their widespread adoption risks conditioning the evolution of their
 72 interactions towards overly homogeneous ones, a phenomenon we denote by *steamrolling*.

73 2.1 Current Embodied Agents are Stagnant

74 Currently, we implicitly assume that there exists a predefined underlying data distribution, from
 75 which we can extract representative examples, to train and evaluate the behavior of embodied agents:
 76 for example, over the sentences people use when feeling happy, or over the possible socially accepted
 77 distances from humans while navigating a crowded room. Furthermore, it is assumed that this data
 78 distribution is static in time. As such, most of the knowledge acquisition and behavior exploration by
 79 the agent happens before it is deployed in a specific environment⁴. This inability to deal with changes
 80 in their own knowledge and their environment leads to their *stagnant* nature.

³Our position builds on past parallels between computers and biological processes [147, 19, 26].

⁴Recent approaches for training embodied agents additionally use fine-tuning to adapt the pretrained behavior of the agent to a specific task. We note that the fine-tuning data distribution is also itself predefined and static.



Figure 1: **Embodied artificial agents must coexist.** Current agents exist in the real-world, leveraging knowledge obtained from large-scale datasets and specific expert-level datasets to interact. We argue that embodied artificial agents must not only adapt to scenarios such as the ones pictured above but participate in their continual evolution. To do so, they must *coexist*, establishing meaningful and reciprocal relationships with the user and its particular environment by leveraging their diverse and situated knowledge.

81 Humans will adapt and change their behavior according to the environment they are situated in,
 82 but also participate in its shaping [33]. A classic example can be found in medical record cards in
 83 hospital beds: Nygren and Henriksson [103] found that the physical properties of the card (e.g., the
 84 handwriting, wear, tear, other marks) were contributing to the physician’s decisions pertaining to both
 85 the patient and the activities surrounding their care. The hospital’s culture and workflows are not
 86 converging to a “fixed” version, rather, they are perpetually evolving as the people, the environment
 87 and their interactions change. This is not only happening on a high functioning level: Vergunst
 88 and Ingold [142] show that even lower level motor skills, such as the way humans walk, are highly
 89 socialized and both culturally and contextually dependent. Therefore, a stagnant agent placed in
 90 this system would not be able to participate in this mutual shaping, as its behavior is a function of
 91 knowledge from a fixed point in time, which can be outdated at deployment time. Even a well-adapted
 92 agent at deployment will drift from the culture as the system evolves.

93 2.2 Current Embodied Agents are Generic

94 Recent advances in machine learning have enabled the extraction of general rules (e.g., grammar
 95 and social norms) from large-scale data to bootstrap the behavior of embodied agents [132, 154].
 96 While learning general rules is valuable, we emphasize the crucial distinction between being generic
 97 and being general. *General* knowledge captures fundamental principles that apply broadly across a
 98 wide range of cases, enabling generalization, a desirable feature in both autonomous systems and
 99 humans ⁵. In contrast, *generic* knowledge is applied across many situations without accounting
 100 for their specific nuances or contextual diversity. By learning generic information from large-scale
 101 datasets, agents reinforce (potentially harmful) biases that exist on such data [109]: for example,
 102 image generation models produce images of white men for the prompt “a software engineer” and
 103 women with darker skin tone for the prompt “a housekeeper” [13]. Current embodied agents, which
 104 often employ such models for interaction purposes, also rely on generic knowledge[7]. Exploiting
 105 only generic knowledge is also inefficient. For example, compare a highly controlled space such as
 106 a factory, where workbenches and machines are specifically configured, to a home or office space.
 107 Each instance of a home or office is unique and contains situated knowledge that is specific to its
 108 configuration and the humans in it [33]. An agent that relies solely on generic knowledge, is at a
 109 clear disadvantage against an agent that also exploits situated knowledge and, just as importantly,
 110 contributes to the ongoing exploration and exploitation of culture and workflow in the space [46].

111 2.3 Current Embodied Agents Will Steamroll

112 When a stagnant and generic agent is placed in a dynamic environment, either the agent will become
 113 obsolete and removed from the system over time ⁶, or the humans will adapt to it. Therefore, in the
 114 latter case, human cultures and workflows will start to converge towards those dictated by the agent,

⁵In Appendix A we discuss the challenges of generalization for current embodied agents.

⁶See Appendix D for the case of the *novelty effect* in the current deployment of social robots.

115 limiting the exploration and discovery of novelty [135]. We denote this phenomenon as *steamrolling*.
 116 Whilst the current prevalence of artificial embodied agents in everyday settings limits the empirical
 117 evidence to support steamrolling, we point to convergent effects in similar socio-technical spaces that
 118 foreshadow steamrolling. Geng and Trotta [44] estimates that 35% of all scientific paper abstracts
 119 in computer science are now written in “LLM-style”. Székely et al. [133] argue that such a style is
 120 spreading from text interactions to more embodied interactions such as speech. Meincke et al. [94]
 121 have shown that in brainstorming sessions, despite having a positive impact on individual creativity,
 122 the pool of LLM-assisted responses exhibits lower diversity. They argue that effective brainstorming
 123 is undermined by the collective use of LLM-tools because instead of creating diverse ideas, similar
 124 thoughts were repeated by multiple participants. In the context of embodied artificial agents, we
 125 expect steamrolling to inhibit divergent behavior, in favor of reinforcing already existing behavior
 126 of both human and agents. We argue that steamrolling will also affect the future capabilities of the
 127 agents we develop: a model trained on a progressively narrower distribution (such as data curated
 128 from its own outputs) suffers from rapid degradation in the quality of its generated output [126].

129 3 Future Embodied Agents Must Coexist

130 Long-term interactions between humans and embodied artificial agents have been extensively studied
 131 by the robotics community [85, 31, 79], focusing on specific properties of the interaction such as
 132 acceptance [31], engagement [115, 84] and disclosure [98, 89]. Here we take a holistic view of the
 133 long-term interactions of embodied agents within a system and provide a general-purpose, formal
 134 definition of coexistence⁷.

Definition: An embodied artificial agent is *coexisting* in a system if it sustains meaningful and reciprocal interactions with humans and their environment over time.

135
 136 Consider a system $S = \{A, H, E\}$ consisting of an embodied agent A_t present in a specific envi-
 137 ronment E_t alongside a human user H_t , at a given time t . There exists a quality function $Q_O(t)$
 138 that overall describes the system and its evolution, measured from the point of view of an observer
 139 $O \in S$. The quality function is influenced by the interactions between the agent, the user and the
 140 environment⁸. We note that the goal of the agent does not necessarily align with this quality function
 141 as it may be independent of its intended task (e.g., a household robot assisting with chores may
 142 perform its tasks efficiently but disrupt the human’s workflow and create frustration).

143 We can define two categories of interactions within this system. A unilateral interaction $X_t \rightarrow Y_t$
 144 occurs if the state of element Y of the system at the next time step $(t + 1)$ is influenced by element
 145 X , while the next state of X remains independent of Y ,

$$Y_{t+1} = f_Y(Y_t, X_t, y_t, x_t), \quad X_{t+1} = f_X(X_t, x_t), \quad (1)$$

146 where f_X, f_Y are unknown and dynamic transition functions, and x_t, y_t are the actions of X and Y
 147 at time t . Similarly a *reciprocal* interaction $X_t \leftrightarrow Y_t$ occurs if the next state of both elements are
 148 mutually influenced,

$$Y_{t+1} = f_Y(Y_t, X_t, y_t, x_t), \quad X_{t+1} = f_X(X_t, Y_t, x_t, y_t). \quad (2)$$

149 Interactions influence the long-term quality of the system, which can be measured after a (system-
 150 dependent) time horizon threshold T_S . We define a *meaningful* interaction as one that, given sufficient
 151 time (i.e., in the long-term), does not decrease the overall quality of the system, as evaluated by all
 152 elements of the interaction, compared to the absence of such interaction. Formally,

$$\exists T_S > t, \forall t' > T_S, \forall O \in \{X, Y\} : Q_O(t' | X_t \rightarrow Y_t) \geq Q_O(t' | \emptyset), \quad (3)$$

153 where $\emptyset$ denotes no interaction and the conditional quality function $Q_O(t' | X_t)$ indicates the value
 154 of the quality function at t' given that process X occurred at $t < t'$. A coexisting agent A^* is
 155 then defined as an agent able to maintain reciprocal and meaningful interactions in the long-term.
 156 Intuitively, this means that, in the long run, the agent benefits the system more than its removal would,

$$\exists T_S > t, \forall t' > T_S, \forall O \in \{A^*, H\} : \quad (4)$$

$$Q_O(t' | A_t^* \leftrightarrow (H_t, E_t), H_t \leftrightarrow E_t) \geq Q_O(t' | H_t \leftrightarrow E_t).$$

⁷In Appendix B we present additional considerations and limitations of our definition of coexistence.

⁸In Appendix C we discuss in depth how to potentially instantiate and measure the quality of the system.

157 3.1 Properties of Coexisting Embodied Agents

158 **Situatedness** A coexisting agent A^* should actively leverage the fact that it is situated within
 159 a specific environment and exploit the unique situated knowledge embedded in the user and their
 160 environment, rather than relying solely on pretrained knowledge. This capability reflects the agent’s
 161 *speciation* to its particular system. Formally, this can be expressed as:

$$\begin{aligned} \exists T_S > t, \forall t' > T_S, \forall O \in \{A^*, H\}, \forall O' \in \{A^*, H'\} : \\ Q_O(t' \mid A_t^* \leftrightarrow (H_t, E_t)) \geq Q_{O'}(t' \mid A_t^* \leftrightarrow (H'_t, E'_t)), \end{aligned} \quad (5)$$

162 where we define a distinct system $S' = \{A^*, E', H'\}$ with its own specific quality function $Q_{O'}(t)$,
 163 but involving the same agent. Note that, contrary to the generic nature of current embodied agents,
 164 we argue that the behavior of coexisting agents should improve the quality of their specific system,
 165 even if the same behavior would result in a overall quality decrease in other distinct systems.

166 **Mutability** A coexisting agent A^* should be capable of continuously adapting its behavior while
 167 also influencing the behavior of other elements within the system. Formally, this adaptability relates
 168 to the concept of reciprocal interactions:

$$\begin{aligned} \exists T_S > t, \forall t' > T_S, \forall O \in \{A^*, H\} : \\ Q_O(t' \mid (H_t, E_t) \leftrightarrow A_t^*) > Q_O(t' \mid (H'_t, E'_t) \rightarrow A_t^*). \end{aligned} \quad (6)$$

169 This condition implies that coexisting agents and humans should be able to mutually shape each other
 170 in ways that enhance the overall quality of the system. In contrast, the stagnant nature of current
 171 embodied systems often requires a unidirectional training of the human user.

172 Importantly, changes in the agent’s behavior do not always lead to an immediate improvement in
 173 system quality and may sometimes have the opposite effect. As discussed in Section 4, coexisting
 174 agents must be capable of generating divergent behavior even within a closed system. This ability is
 175 crucial for the long-term success of the system as it enables the exploration of alternative solutions,
 176 not only in the agent’s behavior but also in how its behavior impacts the other elements of the system.

177 4 Coexistence Elsewhere

178 Beyond the formalism of coexistence, the question remains of how to instantiate embodied agents
 179 with the ability to coexist in the real-world⁹. To address this challenge, we take an interdisciplinary
 180 approach to understand how humans and non-human organisms foster entities that coexist at different
 181 levels of organization: from the processes of biology to the methods of design theory. We explore
 182 research in these fields that highlight the value of mutability and situatedness in fostering coexistence.

183 4.1 Coexistence in Biology

184 Biological systems offer a unique perspective on coexistence, showing how living organisms evolve,
 185 adapt, and sustain themselves in their own environments. Unlike current embodied agents, which
 186 assume that all necessary knowledge can be extracted from data and encoded, biology balances
 187 encoded information with meaningful interactions with the physical world to shape adaptation,
 188 survival and purpose. In this section, we present examples from genetics and developmental biology
 189 that explore how biology navigates this balance.

190 **Not everything is in the genome** Underlying the majority of machine learning models is the
 191 assumption that all necessary knowledge to act/decide optimally can be extracted from data and
 192 subsequently exploited. However, biology provides a perspective shift in regards to the nature and
 193 role of data in the evolution of agents. To illustrate how encoded information is only one part of
 194 what shapes biological organisms, we turn to the Human Genome Project [27]. When this project
 195 successfully sequenced the entire human genome it was widely believed that the genome could define
 196 what humans are, an “instruction book for life”. However as Ball [9] explains, the project instead
 197 marked the beginning of a paradigm shift in biology that de-throned the genome as an encrypted

⁹In Appendix D, we examine if current agents already coexist, providing examples on why they fall short.

198 source of life's secrets. Instead it was shown that an organism is not only defined by the genome but
199 also by principles of self-organization that are enacted by being situated in the physical world [9].

200 A striking example of this new reality can be seen in developmental biology, where the number,
201 thickness, and size of a rodent's digits were not found to be encoded in the genome. Instead, a timing
202 of particular proteins (namely BMP, SOX9 and WNT) that disperse in physical space determines the
203 number of digits and the space between them. Raspopovic et al. [116] discovered that they could
204 manipulate the activity of these proteins and could thus influence the number of digits formed and
205 their thickness. This example shows how the characteristics of the physical world play a role in
206 defining information and intelligence, providing an extremely efficient way of acting in the world [9].

207 **Biology is not an optimizer** Leveraging the physical world is not only about converging on
208 optimally efficient solutions but also about diverging from locally competitive landscapes. It is a
209 common misconception that biology is an optimizer. As Stanley and Lehman [128] write: "Early
210 evolutionists believed, and indeed many non-experts still believe, that evolution is progressive, moving
211 towards some sort of objective perfection, a kind of search for the über organism". In fact, "most
212 evolutionary changes at the molecular level [DNA] are caused not by Darwinian selection but by
213 random *genetic drift* of mutated genes that are selectively neutral" [149].

214 As an example, consider the protein HSP90, where HSP denotes for "heat shock protein". HSP90
215 was discovered to have a kind of plasticity modulation effect on the body plans of the common fruit
216 fly. In warmer conditions, this protein enables more variation in the morphology of the fruit flies, in
217 places such as its abdomen, bristles, eyes, legs, thorax and wings [120]. In addition, these traits were
218 able to be passed down immediately to the next generation [149]. It is argued that processes like
219 the ones observed here played a large part in periods of intense diversification in living organisms
220 during the Cambrian explosion [9]. This alludes to the idea that evolution, whilst highly divergent, is
221 both *bound* and *liberated* by the laws of nature: by using existing building blocks in creative ways,
222 it is able to keep a tension between convergence and divergence [45], conditioning and stimulating
223 exploration and exploitation of novel solutions within its own laws.

224 4.2 Coexistence in Design

225 We have seen how biological organisms exploit being situated in the world to balance convergence
226 and divergence in order to foster coexistence in their physical setting. However, how a human could
227 instantiate a similar process, with their plans, goals, morals, and aesthetics is still unclear. The
228 answer lies in the divergent and convergent *processes* of design which cause an individual to engage
229 reciprocally with technology and its environment, as highlighted in Figure 2.

230 **The double diamond** The design process often converges to a design outcome, due to performance
231 specifications [29], or intended functions or styles [119]. In order to deliver an outcome, methods and
232 heuristics exist within each design discipline [136, 29]. But beneath these formalizations lies a practice
233 that is tacit and with an improvisational dimension. This dimension is not only a function of expert
234 knowledge from formal education (industrial, mechanical, electrical, graphical, architectural, etc.),
235 but a craft-like knowledge of their materials, and a situated understanding of how to use them, built up
236 over years of experience [124]. This process is popularly characterized by the UK Design Council's
237 double diamond [125], highlighted in Figure 2a. Initially when a designer receives a specification,
238 they begin to explore *divergently* how to think about the problem: this involves reasoning about the
239 materials, context, people, social structures, and policy context of the request [136]. Subsequently,
240 they begin to *converge* on a more concrete definition of the problem and present it to the stakeholders
241 involved. At this moment, all stakeholders *diverge* again, exploring various designs without limits as
242 they explore the potential solution space. Finally, the designer converges on a solution, synthesizing
243 all that they have learned to present a design that is on time, budget, and to specification. The double
244 diamond merges a designer's expertise with their situated knowledge and experience.

245 The outcome-centered perspective inherent in the double diamond brings with it the notion that
246 a design should be finished and then deployed in its "finished state" [137, 117]. Here we find an
247 interesting bridge to current embodied artificial agents: they too pass through a phase of training
248 and are only subsequently deployed when they have reached a pre-defined threshold of performance.
249 In interaction design, this perspective limits a finished design to its intended function. Despite the
250 efforts of human factors, user-centered design and participatory design methods [125], ethnographic

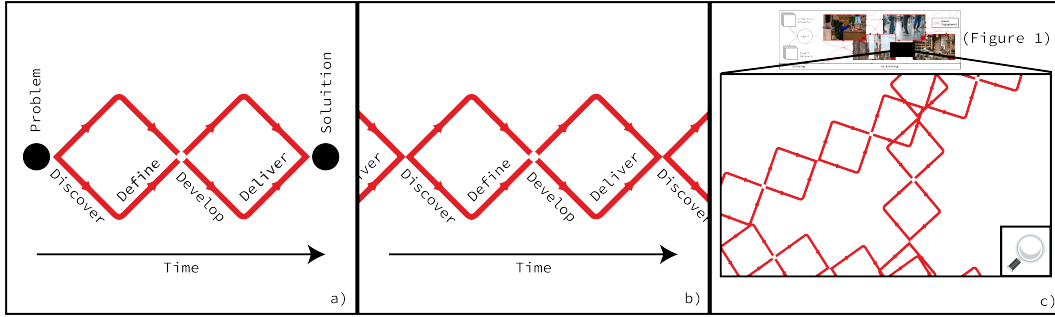


Figure 2: **The evolution of coexisting embodied agents:** a) The double diamond process, with its distinct problem/solution-focused beginning and end; b) Removing the head and tail off the double diamond reveals a continuous and reflective engagement with technology as demonstrated by the field of research through design; c) Revisiting Figure 1, by involving end-users as design researchers, they are encouraged to draw from their experience to integrate technology into existing contexts and to actively shape and explore new ones.

studies often reveal the user to be constantly spending time and creative energy to configure these finished designs and their intended functions into their own lives [33, 129, 36, 102]. This has lead to the increasingly blurred line between what constitutes a designer and a user of technology [117].

Research through design (the continuous double diamond) The field of human-computer interaction (HCI) has seen in the last two decades the rise of *research through design* (RtD) [69, 43] which supports the notion that design is never finished. It is commonly framed as “an active process of ideating, iterating, and critiquing potential solutions, design researchers continually reframe the problem as they attempt to make the right thing” [160]. RtD can be understood as a continuous double diamond (see Figure 2b), with its tail (problem) and head (solution) lopped off. The design process then becomes reflective: where the morals, lived experience, and aesthetic preferences of the designer¹⁰ can inform their professional training [76], leading to completely new (divergent) ways of interacting with technology [11], or familiar (convergent) twists on existing ones [104]. This kind of continuous design has been termed “drifting” by Krogh et al. [72] and bears a striking, functional resemblance the genetic drift discussed in Section 4.1.

4.3 From Elsewhere to Embodied Agents

By exploring coexistence in biology, we have shown that living organisms leverage the physical world to offload the need for encoding all necessary information for survival and action, while also enabling diverse and adaptable behaviors. By exploring coexistence in design, we have highlighted RtD as a promising approach to balance convergence and divergence in the interaction between humans and technology. A common thread between these explorations is the notion of drifting, a process that allows systems at different levels of organization (from microorganisms to human designers) to cope with changing objectives (mutability) across different environments (situatedness) over time. In the next section we extend these ideas to embodied agents: leveraging the situated knowledge in the environment and in the human user enables embodied agents to successfully change, evolve and interact in a meaningful way within their specific environments.

5 Towards Coexisting Embodied Agents

We have seen how both human and non-human organisms evolve and coexist within their own environments. What can the AI community learn from these processes? **This section outlines six key research directions toward fostering coexistence and developing coexisting agents**, focusing on the *principles* that shape coexistence (1, 2), the *hardware* that supports it (3, 4) and the *methods* that enable it (5, 6). Finally, we address ethical considerations of coexistence.

1) Foster coexistence by embracing open-endedness Hughes et al. [58] argues for open-endedness to design continuously evolving agents, defining it as a property of systems that produce novel

¹⁰As a “first person method” [91] RtD can trace its theoretical foundations to the theories of embodiment [81].

and learnable artifacts from the perspective of an observer. We propose that open-endedness is an essential principle for the design of coexisting agents. They should be able to continuously evolve with their environment, changing with it and contributing to its change. This means that both the agent as well as its objectives are non-static, creating the need for open-endedness. We agree that open-endedness is essential to achieve coexisting agents, and highlight the shared importance of the observer’s perspective between open-endedness and RtD. We see the role of the observer as a driver of continuous change and exploration, not just a creative optimizer for a given task.

2) Foster coexistence by embracing the user as the designer Often the user is seen as someone who should not have to deal with the complexities that arise from interacting with technology [102]. In RtD, this perspective is rejected in favor of seeing the user as someone who has situated knowledge, or is a *connoisseur* of their situation [160, 91]. Situated knowledge includes tacit, institutional, craft or social knowledge, and can help mediate an agent’s purpose or behavior in an environment. We argue that this perspective is essential to coexistence and should guide the development of embodied agents. In Appendix F we provide some examples that demonstrate the potential of this principle.

3) Foster coexistence in the space around the agent Consider an agent using an inside-out navigation system, (e.g., SLAM [35]) which is inherently prone to drift. If an outside-in navigation system is instead used (where the agent navigates relative to a set of beacons), the agent can be designed such that the situated human can configure the placement of the beacons. Whilst this sounds like a poorly designed system that requires constant maintenance¹¹ research on AI education has favored this more active and experiential approach, as it fosters a kind of tacit understanding of the capabilities and limitations of the system [41, 150, 63]. The situated knowledge gained from this approach can help users coexist with agents in a specific environment.

4) Foster coexistence within the morphology of the agent Evolutionary robotics has demonstrated that by changing the morphology of an artificial agent, you change their capabilities and limitations [110]. Additionally, advancements in manufacturing technology are rapidly expanding the potential forms an agent could take [70]. This concept has been explored in the context of human-drone interaction. La Delfa et al. [77] gave users a drone that could initially only hover in place. By moving with the drone, the users were able to selectively expand its perceptive field. As the field grew in size, unique patterns of interaction emerged based on its the shape and size. The mutability of the drone’s sensory field allowed for a meaningful relationship to evolve.

5) Foster coexistence by using foundation models as external components Recent methods have used foundation models or composite systems that incorporate foundation models to generate agent behavior [18]. While using these models directly as policies is not sufficient for coexisting agents, foundation models still have valuable properties that can be leveraged (even if these are currently prone to hallucinations [88, 159]): they can act as an external storage of *generic* knowledge that an agent could query for bootstrapping purposes *without replacing situated knowledge*. This external knowledge base could help decrease the memory and computation requirements to build embodied agents [40]. Additionally, foundation models could serve as external teachers to agents to bootstrap their performance [151] and guide exploration [74] *without replacing situated exploration*. While we understand these models can also be used for multimodal perception and reasoning, we highlight the risk of embedding such internal components of embodied agents with generic and stagnant knowledge and encourage researchers to consider using the real-world as “its own best model”[110].

6) Foster coexistence by learning and evolving with humans as we go In their current form, even common learning approaches designed to overcome the assumption of a static optimization problem (e.g., online reinforcement learning, meta-learning, and continual learning) are insufficient to foster coexistence¹². To enable mutability and speciation, instead, we advocate for human-in-the-loop learning with evolutionary algorithms [86, 22]. Evolutionary algorithms [8, 87] can maintain diverse candidate solutions throughout the (continuous) learning process, allowing agents to execute multimodal behavior, both divergent and convergent [97]. When combined with interactive learning paradigms [155, 96], such as by using preferences or demonstrations, these evolutionary processes

¹¹As a comparison, we would like to highlight the resources required to create and curate large-scale datasets.

¹²For an extended argument on why this is the case we refer the reader to Appendix E

334 can also be progressively shaped through meaningful interactions with the human, allowing the agent
335 to deal with evolving goals and expectations.

336 5.1 Should We Foster Coexistence?

337 Coexistence gives users the ability to shape and be shaped by embodied agents, carrying the inherent
338 risk of manipulation of the agent’s behavior by malicious users [100] and vice versa. However, when
339 users are given the responsibility to shape the agents in their environment, we enable them to do so
340 in their own particular way, resulting in a *heterogeneous* population of bespoke agents. In contrast,
341 Székely et al. [133] and Bongard [15] warn of the risks of malicious manipulation of human users
342 using homogeneous artificial intelligence agents at internet-scale. We still highlight the importance
343 of developing agents that have the ability to recognize harmful behavior and respond in a manner that
344 upholds safety, fairness, and accountability.

345 The heterogeneity inherent to coexistence, requires a continuous effort and can be a slower process
346 compared to unidirectional alignment. It also involves high level of ambiguity (e.g., the exploration of
347 a solution space without a set objective) and an active, reflective engagement on behalf of the end-user
348 (e.g., the evaluation of said exploration). Similarly, local and diverse AI collectives have been shown
349 to exhibit strong innovative capabilities and pro-social behavior [80]. Still, they should not be treated
350 as the same kind of relationships: we highlight the need for ethical and legal frameworks that elevate
351 human well-being and safety above that of artificial agents.

352 6 Alternative Views to Coexistence

353 **AGI/ASI vs. coexistence** While coexistence is a goal and property in itself, other positions argue
354 for different goals and capabilities of long-term interactive artificial agents within our societies. Paolo
355 et al. [108] argue in favor of attempting to achieve artificial general intelligence (AGI), describing
356 the goal as “creat[ing] intelligence that either parallels or exceeds human abilities”. They state that
357 embodiment and situated intelligence are essential conditions for achieving AGI. Similarly, Hughes
358 et al. [58], argue in favor of artificial superhuman intelligence (ASI) and propose open-endedness
359 as a prerequisite to ASI. Whilst we share an understanding of the importance of embodiment and
360 open-endedness, neither position requires mutual co-shaping for the widespread use of artificial
361 agents in human society. Despite its risks [99, 93], AGI and ASI proponents point to the accelerated
362 progress and benefit for humanity driven by a single superior intelligence. Instead, we believe that
363 through the increase in diversity, coexistence aims for something more beneficial and robust: we
364 place meaningful and reciprocal interactions with humans at the center of our proposal.

365 **Unilateral alignment vs. coexistence** Yang et al. [152] state that “unified alignment between
366 agents, humans and their environment” is key to the success of agents in real-world applications. They
367 propose that agents not only align with human users, but also with the environment and the agent’s
368 own constraints. Furthermore, they highlight the difficulty of discovering human intentions due to
369 partial observability, temporality and stochasticity. Although they discuss the need for agents that can
370 align with evolving preferences, a process they denote as continual alignment, they still assume that
371 preferences are something that is known by the human a priori. They write: “the tasks assigned by
372 humans can be viewed as the initial inputs to the working system (especially to the agents), which
373 reflects the underlying goals and human intentions”. We instead believe that the human’s goals are
374 formed through interacting with the agent.

375 7 Conclusion

376 In this paper, we have argued that the current paradigm for designing embodied artificial agents is
377 fundamentally ill-suited for long-term, in-the-wild human interaction. We proposed *coexistence* as a
378 new paradigm for the design of embodied agents that emphasizes meaningful, reciprocal interactions
379 sustained over time. Drawing from biology and design, we showed how human and non-human
380 organisms leverage the physical world in convergent and divergent ways. We outlined key research
381 directions for coexisting agents, emphasizing open-ended, human-in-the-loop learning and the user’s
382 role in shaping both behavior and morphology. We envision a future where artificial agents do not
383 just exist but *coexist*, actively shaping and adapting with humans and their environments.

References

- [1] Abel, D., Barreto, A., Van Roy, B., Precup, D., van Hasselt, H. P., and Singh, S. (2023). A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36:50377–50407.
- [2] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Al-tenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- [3] Admoni, H. and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63.
- [4] Alonso, E., Jelley, A., Micheli, V., Kanervisto, A., Storkey, A., Pearce, T., and Fleuret, F. (2024). Diffusion for world modeling: Visual details matter in atari. *arXiv preprint arXiv:2405.12399*.
- [5] Aly, A., Griffiths, S., and Stramandinoli, F. (2017). Metrics and benchmarks in human-robot interaction: Recent advances in cognitive robotics. *Cognitive Systems Research*, 43:313–323.
- [6] Auger, J. (2022). Seven Observations, or Why Domestic Robots are Struggling to Enter the Habitats of Everyday Life. In *Meaningful Futures with Robots*. Chapman and Hall/CRC.
- [7] Azeem, R., Hundt, A., Mansouri, M., and Brandão, M. (2024). Llm-driven robots risk enacting discrimination, violence, and unlawful actions.
- [8] Bäck, T. and Schwefel, H.-P. (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary computation*, 1(1):1–23.
- [9] Ball, P. (2023). *How Life Works: A User’s Guide to the New Biology*. Pan Macmillan.
- [10] Belkaid, M., Kompatsiari, K., De Tommaso, D., Zablit, I., and Wykowska, A. (2021). Mutual gaze with a robot affects human neural activity and delays decision-making processes. *Science Robotics*, 6(58):eabc5044.
- [11] Bewley, H. and Boer, L. (2018). Designing blo-nut: Design principles, choreography and otherness in an expressive social robot. In *Proceedings of the 2018 Designing Interactive Systems Conference, DIS ’18*, page 1069–1080, New York, NY, USA. Association for Computing Machinery.
- [12] Bharadhwaj, H. (2024). Position: scaling simulation is neither necessary nor sufficient for in-the-wild robot manipulation. In *Forty-first International Conference on Machine Learning*.
- [13] Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’23*, pages 1493–1504. Association for Computing Machinery.
- [14] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. (2024). π_0 : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.
- [15] Bongard, J. (2024). The tip and the iceberg: Deep learning and embodiment. Invited talk at CVPR 2024, Summit Flex Hall ABC. Accessed: 2025-05-21.
- [16] Bowling, M. and Elelimy, E. (2025). Rethinking the foundations for continual reinforcement learning. *arXiv preprint arXiv:2504.08161*.
- [17] Breazeal, C., Dautenhahn, K., and Kanda, T. (2016). Social robotics. *Springer handbook of robotics*, pages 1935–1972.

- [18] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.-W. E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*.
- [19] Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159.
- [20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- [21] Bruce, J., Dennis, M. D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al. (2024). Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*.
- [22] Bryden, K. (2006). Using a human-in-the-loop evolutionary algorithm to create data-driven music. In *2006 IEEE International Conference on Evolutionary Computation*, pages 2065–2071. IEEE.
- [23] Chen, M., Nikolaidis, S., Soh, H., Hsu, D., and Srinivasa, S. (2018). Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*, pages 307–315.
- [24] Chen, N., Liu, X., Zhai, Y., and Hu, X. (2023). Development and validation of a robot social presence measurement dimension scale. *Scientific Reports*, 13(1):2911.
- [25] Chen, X., Hu, J., Jin, C., Li, L., and Wang, L. (2021). Understanding domain randomization for sim-to-real transfer. *arXiv preprint arXiv:2110.03239*.
- [26] Clark, A. (2001). *Mindware : an introduction to the philosophy of cognitive science*. New York : Oxford University Press.
- [27] Collins, F. S. and Fink, L. (1995). The human genome project. *Alcohol Health and Research World*, 19(3):190–195.
- [28] Correia, F., Petisca, S., Alves-Oliveira, P., Ribeiro, T., Melo, F. S., and Paiva, A. (2019). “i choose... you!” membership preferences in human–robot teams. *Autonomous Robots*, 43:359–373.
- [29] Cross, N. (2000). *Engineering design methods: strategies for product design*. Wiley, Chichester ; New York, 3rd ed edition.
- [30] Darlow, L., Regan, C., Risi, S., Seely, J., and Jones, L. (2025). Continuous thought machines.
- [31] de Graaf, M. M., Allouch, S. B., and van Dijk, J. A. (2016). Long-term acceptance of social robots in domestic environments: insights from a user’s perspective. In *2016 AAAI spring symposium series*.
- [32] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Hounsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale.
- [33] Dourish, P. (2001). *Where the Action Is: The Foundations of Embodied Interaction*. The MIT Press.
- [34] Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.
- [35] Durrant-Whyte, H. and Bailey, T. (2006). Simultaneous localization and mapping: part i. *IEEE robotics & automation magazine*, 13(2):99–110.

- [36] Dörrenbächer, J., Hassenzahl, M., Neuhaus, R., and Ringfort-Felner, R. (2022). Towards Designing Meaningful Relationships with Robots. In *Meaningful Futures with Robots—Designing a New Coexistence*, pages 3–29. Chapman and Hall/CRC, Boca Raton, 1 edition.
- [37] Esterwood, C., Essenmacher, K., Yang, H., Zeng, F., and Robert, L. P. (2021). A meta-analysis of human personality and robot acceptance in human-robot interaction. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–18.
- [38] Fink, J., Bauwens, V., Kaplan, F., and Dillenbourg, P. (2013). Living with a vacuum cleaning robot. *International Journal of Social Robotics*, 5(3):389–408.
- [39] Finn, C., Abbeel, P., and Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.
- [40] Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K., et al. (2023). Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, page 02783649241281508.
- [41] Flechtner, R. and Stankowski, A. (2023). Ai is not a wildcard: Challenges for integrating ai into the design curriculum. In *Proceedings of the 5th Annual Symposium on HCI Education*, EduCHI ’23, page 72–77, New York, NY, USA. Association for Computing Machinery.
- [42] Frauenberger, C. (2019). Entanglement hci the next wave? *ACM Trans. Comput.-Hum. Interact.*, 27(1).
- [43] Frayling, C. and Royal College of Art. (1993). *Research in art and design*. Number vol. 1, no. 1 in Royal College of Art research papers. Royal College of Art, London. OCLC: 48866129.
- [44] Geng, M. and Trotta, R. (2024). Is chatgpt transforming academics’ writing style? *arXiv preprint arXiv:2404.08627*.
- [45] Gerhart, J. and Kirschner, M. (2007). The theory of facilitated variation. *Proceedings of the National Academy of Sciences of the United States of America*, 104(Suppl 1):8582–8589.
- [46] Gillet, S., Vázquez, M., Andrist, S., Leite, I., and Sebo, S. (2024). Interaction-shaping robotics: Robots that influence interactions between other agents. 13(1):12:1–12:23.
- [47] Glickman, M. and Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. pages 1–15. Publisher: Nature Publishing Group.
- [48] Gombolay, M. (2024). Human-robot alignment through interactivity and interpretability: Don’t assume a “spherical human”. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8523–8528.
- [49] Gronauer, S. and Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943.
- [50] Guo, Y. and Yang, X. J. (2021). Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 13(8):1899–1909.
- [51] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527.
- [52] Hara, S., Tsuchiya, M., and Kimura, T. (2021). Robust Control of Automatic Low-Speed Driving Motorcycle MOTOROiD. In *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pages 645–646, Kyoto, Japan. IEEE.
- [53] Hebart, M. N., Contier, O., Teichmann, L., Rockter, A. H., Zheng, C. Y., Kidder, A., Corriveau, A., Vaziri-Pashkam, M., and Baker, C. I. (2023). Things-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *Elife*, 12:e82580.

- [54] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.
- [55] Hoffman, G. (2019). Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218.
- [56] Hoffman, G. and Breazeal, C. (2007). Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 1–8.
- [57] Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169.
- [58] Hughes, E., Dennis, M. D., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., and Rocktäschel, T. (2024). Position: Open-endedness is essential for artificial superhuman intelligence. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20597–20616. PMLR.
- [59] Jordan, S. M., White, A., Da Silva, B. C., White, M., and Thomas, P. S. (2024). Position: Benchmarking is limited in reinforcement learning research. *arXiv preprint arXiv:2406.16241*.
- [60] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.
- [61] Kamino, W., Jung, M. F., and Sabanović, S. (2024). Constructing a social life with robots: Shifting away from design patterns towards interaction ritual chains. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, HRI '24*, page 343–351, New York, NY, USA. Association for Computing Machinery.
- [62] Kanda, T., Sato, R., Saiwaki, N., and Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on robotics*, 23(5):962–971.
- [63] Kazemitabaar, M., Hou, X., Henley, A., Ericson, B. J., Weintrop, D., and Grossman, T. (2024). How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research, Koli Calling '23*, New York, NY, USA. Association for Computing Machinery.
- [64] Khanna, P., Yadollahi, E., Björkman, M., Leite, I., and Smith, C. (2023). Effects of explanation strategies to resolve failures in human-robot collaboration. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1829–1836.
- [65] Khavas, Z. R. (2021). A review on trust in human-robot interaction. *arXiv preprint arXiv:2105.10045*.
- [66] Kidd, C. D. and Breazeal, C. (2008). Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3230–3235. IEEE.
- [67] Kim, G., Xiao, C., Konishi, T., and Liu, B. (2023). Learnability and algorithm for continual learning. In *International Conference on Machine Learning*, pages 16877–16896. PMLR.
- [68] Kompatsiari, K., Tikhanoff, V., Ciardo, F., Metta, G., and Wykowska, A. (2017). The importance of mutual gaze in human-robot interaction. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, pages 443–452. Springer.
- [69] Koskinen, I. K. (2011). *Design research through practice: from the lab, field, and showroom*. Morgan Kaufmann/Elsevier, Waltham, MA.
- [70] Kriegman, S. (2020). Design for an Increasingly Protean Machine. *Graduate College Dissertations and Theses*.

- [71] Kriegman, S., Walker, S., Shah, D. S., Levin, M., Kramer-Bottiglio, R., and Bongard, J. (2019). Automated shapeshifting for function recovery in damaged robots. In *Proceedings of Robotics: Science and Systems*, Freiburg/Breisgau, Germany.
- [72] Krogh, P. G., Markussen, T., and Bang, A. L. (2015a). Ways of drifting—five methods of experimentation in research through design. In Chakrabarti, A., editor, *ICoRD'15 – Research into Design Across Boundaries Volume 1*, pages 39–50, New Delhi. Springer India.
- [73] Krogh, P. G., Markussen, T., and Bang, A. L. (2015b). Ways of drifting—five methods of experimentation in research through design. In Chakrabarti, A., editor, *ICoRD'15 – Research into Design Across Boundaries Volume 1*, pages 39–50, New Delhi. Springer India.
- [74] Kumar, A., Lu, C., Kirsch, L., Tang, Y., Stanley, K. O., Isola, P., and Ha, D. (2024). Automating the search for artificial life with foundation models. *arXiv preprint arXiv:2412.17799*.
- [75] La Delfa, J. (2023). *Cultivating Mechanical Sympathy : Making meaning with ambiguous machines*. PhD thesis, KTH, Media Technology and Interaction Design, MID. QC 20230927.
- [76] La Delfa, J., Baytaş, M. A., Luke, E., Koder, B., and Mueller, F. F. (2020). Designing drone chi: Unpacking the thinking and making of somaesthetic human-drone interaction. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, DIS '20, page 575–586, New York, NY, USA. Association for Computing Machinery.
- [77] La Delfa, J., Garrett, R., Lampinen, A., and Höök, K. (2024a). How to train your drone: Exploring the umwelt as a design metaphor for human-drone interaction. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, page 2987–3001, New York, NY, USA. Association for Computing Machinery.
- [78] La Delfa, J., Garrett, R., Lampinen, A., and Höök, K. (2024b). Articulating mechanical sympathy for somaesthetic human–machine relations. In *Proceedings of the 2024 ACM Conference on Designing Interactive Systems*, pages 1–18.
- [79] Laban, G., Kappas, A., Morrison, V., and Cross, E. S. (2024). Building long-term human–robot relationships: Examining disclosure, perception and well-being across time. *International Journal of Social Robotics*, 16(5):1–27.
- [80] Lai, S., Potter, Y., Kim, J., Zhuang, R., Song, D., and Evans, J. (2024). Position: Evolving AI collectives enhance human diversity and enable self-regulation. In *Forty-first International Conference on Machine Learning*.
- [81] Lakoff, G. and Johnson, M. (1985). *Metaphors we live by*. Univ. of Chicago Press, Chicago, Ill., 5. [dr.] edition.
- [82] Lee, J. J., Sha, F., and Breazeal, C. (2019). A bayesian theory of mind approach to nonverbal communication. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 487–496. IEEE.
- [83] Lehman, J., Meyerson, E., El-Gaaly, T., Stanley, K. O., and Ziyadeh, T. (2025). Evolution and the knightian blindspot of machine learning. *arXiv preprint arXiv:2501.13075*.
- [84] Leite, I., Castellano, G., Pereira, A., Martinho, C., and Paiva, A. (2014). Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children. *International Journal of Social Robotics*, 6:329–341.
- [85] Leite, I., Martinho, C., and Paiva, A. (2013). Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, 5:291–308.
- [86] Li, K. and Guo, H. (2024). Human-in-the-loop policy optimization for preference-based multi-objective reinforcement learning. *arXiv preprint arXiv:2401.02160*.
- [87] Li, N., Ma, L., Yu, G., Xue, B., Zhang, M., and Jin, Y. (2023a). Survey on evolutionary deep learning: Principles, algorithms, applications, and open issues. *ACM Computing Surveys*, 56(2):1–34.

- [88] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. (2023b). Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.
- [89] Ligthart, M., Neerincx, M. A., and Hindriks, K. V. (2019). Getting acquainted for a long-term child-robot interaction. In *International Conference on Social Robotics*, pages 423–433. Springer.
- [90] Lim, S. H., Xu, H., and Mannor, S. (2013). Reinforcement learning in robust markov decision processes. *Advances in Neural Information Processing Systems*, 26.
- [91] Loke, L. and Schiphorst, T. (2018). The somatic turn in human-computer interaction. *Interactions*, 25(5):54–5863.
- [92] Lu, H., Yang, G., Fei, N., Huo, Y., Lu, Z., Luo, P., and Ding, M. (2023). Vdt: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311*.
- [93] McLean, S., Read, G. J. M., Thompson, J., Baber, C., Stanton, N. A., and Salmon, P. M. (2023). The risks associated with artificial general intelligence: A systematic review. 35(5):649–663. Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/0952813X.2021.1964003>.
- [94] Meincke, L., Nave, G., and Terwiesch, C. (2025). Chatgpt decreases idea diversity in brainstorming. *Nature human behaviour*, pages 1–3.
- [95] Mesnard, T., Weber, T., Viola, F., Thakoor, S., Saade, A., Harutyunyan, A., Dabney, W., Stepleton, T. S., Heess, N., Guez, A., et al. (2021). Counterfactual credit assignment in model-free reinforcement learning. In *International Conference on Machine Learning*, pages 7654–7664. PMLR.
- [96] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.
- [97] Mouret, J. and Clune, J. (2015). Illuminating search spaces by mapping elites. *CoRR*, abs/1504.04909.
- [98] Naneva, S., Sarda Gou, M., Webb, T. L., and Prescott, T. J. (2020). A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics*, 12(6):1179–1201.
- [99] Naudé, W. and Dimitri, N. (2020). The race for an artificial general intelligence: implications for public policy. 35(2):367–379.
- [100] Neff, G. (2016). Talking to bots: Symbiotic agency and the case of tay. *International journal of Communication*.
- [101] Nikolaidis, S., Hsu, D., and Srinivasa, S. (2017). Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618–634.
- [102] Norman, D. A. (2010). *Living with Complexity*. MIT Press, Cambridge, MA.
- [103] Nygren, E. and Henriksson, P. (1992). Reading the medical record. I. Analysis of physicians’ ways of reading the medical record. *Computer Methods and Programs in Biomedicine*, 39(1-2):1–12.
- [104] Odom, W., Wakkary, R., Hol, J., Naus, B., Verburg, P., Amram, T., and Chen, A. Y. S. (2019). Investigating slowness as a frame to design longer-term experiences with personal data: A field study of olly. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–16, New York, NY, USA. Association for Computing Machinery.
- [105] Oertel, C., Castellano, G., Chetouani, M., Nasir, J., Obaid, M., Pelachaud, C., and Peters, C. (2020). Engagement in human-agent interaction: An overview. *Frontiers in Robotics and AI*, 7:92.

- [106] O’Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., et al. (2023). Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*.
- [107] Paolo, G., Gonzalez-Billandon, J., and Kégl, B. (2024a). Position: A call for embodied ai. In *Forty-first International Conference on Machine Learning*.
- [108] Paolo, G., Gonzalez-Billandon, J., and Kégl, B. (2024b). Position: A call for embodied AI. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F., editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39493–39508. PMLR.
- [109] Parreira, M. T., Gillet, S., Winkle, K., and Leite, I. (2023). How did we miss this? a case study on unintended biases in robot social behavior. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI ’23*, page 11–20, New York, NY, USA. Association for Computing Machinery.
- [110] Pfeifer, R. and Bongard, J. (2006). *How the Body Shapes the Way We Think: A New View of Intelligence*. A Bradford Book. MIT Press, Cambridge, Massachusetts.
- [111] Pfeifer, R., Iida, F., and Bongard, J. (2005). New robotics: Design principles for intelligent systems. *Artificial Life*, 11(1-2):99–120.
- [112] Pignatelli, E., Ferret, J., Geist, M., Mesnard, T., van Hasselt, H., Pietquin, O., and Toni, L. (2023). A survey of temporal credit assignment in deep reinforcement learning. *arXiv preprint arXiv:2312.01072*.
- [113] Platanios, E. A., Saparov, A., and Mitchell, T. (2020). Jelly bean world: A testbed for never-ending learning. *arXiv preprint arXiv:2002.06306*.
- [114] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- [115] Rakhymbayeva, N., Amirova, A., and Sandygulova, A. (2021). A long-term engagement with a social robot for autism therapy. *Frontiers in Robotics and AI*, 8:669972.
- [116] Raspopovic, J., Marcon, L., Russo, L., and Sharpe, J. (2014). Digit patterning is controlled by a bmp-sox9-wnt turing network modulated by morphogen gradients. *Science*, 345:566 – 570.
- [117] Redström, J. (2017). *Making design theory*. MIT Press, Cambridge, Massachusetts.
- [118] Reimann, M., van de Graaf, J., van Gulik, N., Van De Sanden, S., Verhagen, T., and Hindriks, K. (2023). Social robots in the wild and the novelty effect. In *International Conference on Social Robotics*, pages 38–48. Springer.
- [119] Rodgers, P. (2011). *Product design*. Portfolio. Laurence King, London.
- [120] Rutherford, S. L. and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. *Nature*, 396(6709):336–342.
- [121] Sandoval, E. B., Brandstatter, J., Yalcin, U., and Bartneck, C. (2021). Robot likeability and reciprocity in human robot interaction: Using ultimatum game to determinate reciprocal likeable robot strategies. *International Journal of Social Robotics*, 13(4):851–862.
- [122] Sandry, E. (2015). Re-evaluating the form and communication of social robots - the benefits of collaborating with machinelike robots. *Int. J. Soc. Robotics*, 7(3):335–346.
- [123] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.
- [124] Schön, D. A. (1983). *The reflective practitioner: how professionals think in action*. Basic Books, New York.

- [125] Sharp, H., Rogers, Y., and Preece, J. (2023). *Interaction design: beyond human-computer interaction*. John Wiley & Sons, Inc, Hoboken, sixth edition edition.
- [126] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. (2024). AI models collapse when trained on recursively generated data. 631(8022):755–759. Publisher: Nature Publishing Group.
- [127] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- [128] Stanley, K. O. and Lehman, J. (2015). *Why greatness cannot be planned: the myth of the objective*. Springer International Publishing, Cham Heidelberg New York Dordrecht London.
- [129] Suchman, L. A. (2006). *Human-machine reconfigurations plans and situated actions*. Cambridge University Press, Cambridge ; New York, 2nd ed. edition. OCLC: 1035691746.
- [130] Sung, J., Grinter, R. E., and Christensen, H. I. (2010). Domestic robot ecology. *International Journal of Social Robotics*, 2(4):417–429.
- [131] Sutton, R. S., Barto, A. G., et al. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- [132] Szot, A., Schwarzer, M., Agrawal, H., Mazouze, B., Metcalf, R., Talbott, W., Mackraz, N., Hjelm, R. D., and Toshev, A. T. (2023). Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*.
- [133] Székely, É., Miniot, J., and Hejná, M. (2025). Will ai shape the way we speak? the emerging sociolinguistic influence of synthetic voices.
- [134] Tae, M. and Lee, J. (2020). The effect of robot’s ice-breaking humor on likeability and future contact intentions. In *Companion of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 462–464.
- [135] Tao, Y., Viberg, O., Baker, R. S., and Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.
- [136] Tomitsch, M., Borthwick, M., Ahmadpour, N., Cooper, C., Frawley, J., Hepburn, L.-A., Kocaballi, A. B., Loke, L., Núñez-Pacheco, C., Straker, K., and Wrigley, C. (2020). *Design. Think. Make. Break. Repeat: a handbook of methods*. BIS, Amsterdam, revised edition edition.
- [137] Tonkinwise, C. (2004). Is Design Finished? Dematerialisation and Changing Things. *Design Philosophy Papers*, 2(3):177–195.
- [138] Tulli, S., Ambrossio, D. A., Najjar, A., and Rodríguez-Lera, F. J. (2019). Great expectations & aborted business initiatives: The paradox of social robot between research and industry. *BNAIC/BENELEARN*, 1.
- [139] Udandara, V., Prabhu, A., Ghosh, A., Sharma, Y., Torr, P., Bibi, A., Albanie, S., and Bethge, M. (2024). No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [140] Vasco, M., Seno, T., Kawamoto, K., Subramanian, K., Wurman, P. R., and Stone, P. (2024). A super-human vision-based reinforcement learning agent for autonomous racing in Gran Turismo. *Reinforcement Learning Journal*, 4:1674–1710.
- [141] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.
- [142] Vergunst, J. L. and Ingold, T., editors (2016). *Ways of Walking: Ethnography and Practice on Foot*. Routledge, London.

- [143] Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354.
- [144] Wang, L., Zhang, X., Su, H., and Zhu, J. (2024). A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [145] Wang, R., Zhang, J., Chen, J., Xu, Y., Li, P., Liu, T., and Wang, H. (2023a). Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11359–11366. IEEE.
- [146] Wang, Z., Du, Y., Zhang, Y., Fang, M., and Huang, B. (2023b). Macca: Offline multi-agent reinforcement learning with causal credit assignment. *arXiv preprint arXiv:2312.03644*.
- [147] Winograd, T. and Flores, F. (1986). *Understanding Computers and Cognition: A New Foundation for Design*. Language and being. Ablex Publishing Corporation.
- [148] Xiang, J., Tao, T., Gu, Y., Shu, T., Wang, Z., Yang, Z., and Hu, Z. (2024). Language models meet world models: Embodied experiences enhance language models. *Advances in neural information processing systems*, 36.
- [149] Yahara, I. (1999). The role of HSP90 in evolution. *Genes to Cells*, 4(7):375–379.
- [150] Yan, D. (2023). Impact of chatgpt on learners in a l2 writing practicum: An exploratory investigation. *Education and Information Technologies*, 28(11):13943–13967.
- [151] Yang, J., Mark, M. S., Vu, B., Sharma, A., Bohg, J., and Finn, C. (2024a). Robot fine-tuning made easy: Pre-training rewards and policies for autonomous real-world reinforcement learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4804–4811.
- [152] Yang, Z., Liu, A., Liu, Z., Liu, K., Xiong, F., Wang, Y., Yang, Z., Hu, Q., Chen, X., Zhang, Z., Luo, F., Guo, Z., Li, P., and Liu, Y. (2024b). Position: Towards unified alignment between agents, humans, and environment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 56251–56275. PMLR. ISSN: 2640-3498.
- [153] Yarats, D., Kostrikov, I., and Fergus, R. (2021). Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International conference on learning representations*.
- [154] Yuan, Y., Tang, K., Shen, J., Zhang, M., and Wang, C. (2024). Measuring social norms of large language models. *arXiv preprint arXiv:2404.02491*.
- [155] Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.
- [156] Zeng, K.-H., Zhang, Z., Ehsani, K., Hendrix, R., Salvador, J., Herrasti, A., Girshick, R., Kembhavi, A., and Weihs, L. (2024). Poliformer: Scaling on-policy rl with transformers results in masterful navigators. *arXiv preprint arXiv:2406.20083*.
- [157] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. (2022). Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.
- [158] Zhang, Y., Beskow, J., and Kjellström, H. (2017). Look but don’t stare: Mutual gaze interaction in social robots. In *Social Robotics: 9th International Conference, ICSR 2017, Tsukuba, Japan, November 22-24, 2017, Proceedings 9*, pages 556–566. Springer.
- [159] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., et al. (2023). Siren’s song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.
- [160] Zimmerman, J., Forlizzi, J., and Evenson, S. (2007). Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’07, page 493–502, New York, NY, USA. Association for Computing Machinery.



Figure 3: Fink et al. [38] and Sung et al. [130] highlight the various ways users modify their behaviors and living spaces to integrate autonomous robot vacuum cleaners into daily life.

A Additional Notes on Coexistence

Interaction with humans in coexistence When discussing coexistence, one might ask about embodied agents that are not directly interacting with humans. Does coexistence apply to them as well? Why would an embodied agent that has no relationship with humans have to be designed with interaction in mind? We take the perspective that embodied agents that were made by humans are *interacting* with humans, even in cases where such interaction is not evident at first glance. Supporting our perspective is Frauenberger’s [42] seminal review article on the shifting theoretical foundations of the field of human-computer interaction. Frauenberger [42] states that there are no clear boundaries that separate humans from technology. Instead, the purpose of a technology (e.g., an embodied agent such as a robot) is constituted by not only its physical self but its environment and the surrounding human and non-human agents. Therefore, the vast network of interconnected technologies that support our every day lives are said to be *entangled*, and expands the notion of interaction beyond dyadic “turn taking”.

Take for example, an autonomous robot vacuum cleaner: from an earlier interaction perspective, its sole purpose is to clean up after humans. From an entanglement perspective, however, a robot vacuum cleaner is not seen as a simple tool for cleaning, but has an effect that goes beyond that of its original purpose. This is demonstrated by Sung et al.’s [130] field work with robot vacuum cleaners. Their work frequently revealed users cooperating with the robot by preparing the space before it was cleaned, modifying their environment to facilitate easy navigation and even socially interacting with it, as highlighted in Figure 3. Therefore, despite most of the robot’s actions taking place in the absence of any human, the effect of these actions impacts other objects in the home (e.g., moving furniture, routing cables), human practices (e.g., the habit of tidying before the robot begins its operation) and the environment itself (e.g., the installation of threshold ramps in doorways). However, we note that the effects visible in Sung et al. [130] do not, by themselves, constitute coexistence, but only entanglement. For example, the subsequent in-the-wild study on robot vacuum cleaners by Fink et al. [38] draws attention to several cases of users unable to adapt to the robot, as their existing environment and practices were not amenable to the inflexible design of the robot. This highlights how, by designing robot vacuum cleaners with a fixed and narrow purpose (i.e., cleaning after the human), we are inhibiting the continuous evolution of its role and effects, and, therefore, coexistence. We refer the reader to Appendix D, Social robots, for a complementary discussion that focuses more on the social aspects of everyday life with robots.

Isn’t coexistence just optimization? With coexistence, we are proposing a form of learning and development that is not as strictly focused on optimization as current approaches. In classic machine learning, a task is defined by an objective function, which captures what the model/agent should do, and then fitting available data to that objective. The agent learns by optimizing towards this objective. In our position, we provide a different perspective on this practice. We broaden the focus

by paying attention to system-wide quality (motivated by contemporary views on what interaction is, see previous paragraph). In order to uphold this quality, it might occasionally be necessary to go against the objective function, or to acknowledge the fact that a *static objective function for long-term interaction may not even exist*. While classic machine learning approaches try to converge to a solution, we emphasize the necessity to convergence *and diverge*. Optimization has a place in coexistence, but it needs to be balanced with continuous exploration. Without it, the ability for humans and artificial agents to conceive novel objectives (read, to change and evolve as a society) will be severely distorted (see Section 2).

Designing frameworks for coexistence Considering the design of frameworks that implement coexistence in practice, we aim to embrace a continuous process that balances divergence and convergence, inspired by biology and design. As we outline in Section 4.2, the process that produces the design itself (the continuous double diamond) will result in coexisting entities. Given the improvisational and situated nature of RtD, any kind of framework that implies a strict procedure would go against the value of coexistence. Guidance on these kinds of processes is given by Pfeifer et al. [111], who define design principles for Intelligent Systems, as well as Krogh et al. [73], who identify and analyze drifting as a property of continuous design processes. La Delfa [75, p.111] outlines how these process can be combined in the context of embodied artificial agents.

The role of generalization in coexistence As discussed in Section 2.2, generalization – the ability to leverage *general* knowledge to act across similar scenarios – is a desirable feature for embodied agents. However, we would like to highlight the challenges and potential pitfalls of achieving generalization at the cost of employing *generic* knowledge. Consider once again, the example presented in Section 2.1: Nygren and Henriksson [103] found that the physical properties of medical cards of the patients (e.g., the handwriting, wear, tear, other marks) were contributing to the decision-making of the physician. Achieving such nuanced care and attention from embodied artificial agents through generalization would be unfeasible, as it would require the collection, and inclusion in the training procedure, of increasingly fine-grained physical and social cue data. A similar observation is made by [139], where the authors reveal that zero-shot generalization of concepts in multimodal models require an exponential increase in progressively more fine-grained data. One could argue, then, that we should instead focus on developing increasingly realistic simulators. However, such option also appears fundamentally unfeasible. Bharadhwaj [12] argues that “even the best simulators cannot match reality”, concluding that “scaling simulation frameworks is unlikely to directly help with these [manipulation] tasks as each of these would require separate nuanced considerations for faithful simulation”, hindering generalization. We highlight that the impossibility of collecting exponentially fine-grained data in the real-world to pretrain large-scale models, alongside the infeasibility of developing faithful simulators of the real-world, further motivates our argument to, instead, develop agents that are able to continuously leverage the physical and social cues *present in their specific environment* to achieve coexistence.

B Additional Notes on the Definition of Coexistence

Nature of Q_S and T_S Like all the elements in the system, the operationalization and interpretation of the quality function Q_S is dynamic (meaning it changes over time) and specific to every system. In Appendix C we discuss some potential proxy metrics for the quality function. The same can be said of the time horizon T_S : each particular system should have, even if implicitly, a specific time horizon to assess the evolution of the system itself.

Assumptions of coexistence For simplification we have implicitly assumed that our system is *closed*, meaning that the quality of the interaction is only influenced by the elements within the system (environment, human and agent). We have also assumed that there is a single human user in the system. However, we can easily extend this to open systems and multiple users by considering the correspondent interaction terms with additional elements external to the system (e.g., external societal rules, other human and agent members of a team), without a significant change on the definition of coexistence.

Formally, consider a system $S = \{A^*, \mathbf{H}, \mathbf{A}, E\}$ where $\mathbf{H}_t$ represents the state at time t of the set of all the humans in the system and $\mathbf{A}_t$ the state of all additional agents in the system. Let $\mathbf{E}_t$ be

the state of the set of external components to the system (e.g., external societal rules or observers imposed on the system). We still formally define a *coexisting* agent A^* as,

$$\begin{aligned} & \exists T_S > t, \forall t' > T_S, \forall O \in \{A^*, \mathbf{H}, \mathbf{A}\} : \\ & Q_O(t' \mid A_t^* \leftrightarrow (\mathbf{H}_t, \mathbf{A}_t, E_t), \mathbf{H}_t \leftrightarrow E_t, \mathbf{H}_t \leftrightarrow \mathbf{A}_t, \mathbf{A}_t \leftrightarrow E_t) \geq \\ & Q_O(t' \mid \mathbf{H}_t \leftrightarrow E_t, \mathbf{H}_t \leftrightarrow \mathbf{A}_t, \mathbf{A}_t \leftrightarrow E_t). \end{aligned} \quad (7)$$

We would like to point out that, despite the fact that the system is open, and influenced by the interactions with $\mathbf{E}$, these external components of the system are not observers of the quality function Q_O of the system. In practice, this means that their assessment of the overall quality of the system does not influence the quality of the system itself, as it is measured only by the elements that belong to the system. As an example, consider a companion robot interacting with a teenager ($\mathbf{H}$). Parents ($\mathbf{E}$) might disapprove of the informal language style used by the robot. However, their external assessment does not directly affect the quality of the system: instead, it is measured exclusively based on the perceptions and experiences of the teenager and the robot. In another instantiation of the system, one that encompasses both the teenager, the robot and the parents, the latter would naturally influence the quality of that system.

The quality of a system is not monotonically increasing We do not expect the quality of the system to be monotonically increasing over time; in fact, *we argue that it should not*. Formally, there may exist time steps $t_1 < t_2$ such that $Q_O(t_2) < Q_O(t_1)$, even in the case of meaningful and reciprocal interactions. The requirement for coexistence concerns the asymptotic behavior of the quality function over a system-dependent horizon $t' > T_S$.

Credit assignment in coexistence A core challenge in evaluating coexistence lies in the credit assignment problem [112], i.e., determining which interactions are responsible for the changes in the overall quality of the system over time. This challenge is exacerbated by the fact that the effects of an interaction may not immediately be observable, as multiple overlapping interactions may contribute to a shared outcome. Adding to these challenges, the quality function itself is not stationary, due to the evolution of the human, agent and their environment.

Assigning credit for the long-term impact of meaningful and reciprocal interactions remains an open research problem. Future work may explore the use of causal inference [95] or multi-agent reinforcement learning techniques [146] to better disentangle the contributions of individual interactions.

C Measuring Coexistence

One fundamental challenge of implementing coexistence for embodied agents is how to measure the quality function Q_S of a system. As briefly mentioned in Section 3, the human-robot/computer interaction community has proposed several (often complementary) self-reported metrics to evaluate the quality of an interaction. For example,

- **Trust** reflects the human belief that the robot will behave reliably, safely, and as expected [65]. In collaborative tasks, the performance and attributes of the agent strongly influence the user’s trust. Trust, in turn, correlates with better team performance and interaction outcomes [51].
- **Engagement** reflects how involved, attentive and interested the human is during the interaction. It captures the degree of active participation in the interaction by the human user (e.g., paying attention to the agent, responding to it, initiating the interaction) [105].
- **Likeability** measures how pleasant, friendly, and likeable the agent is to the human user during the interaction. Likeability can emerge from the personality of the agent, its appearance or its behavior [121, 134, 28].
- **Social Presence** relates to the feeling that the agent is a social entity present in the system, as opposed to being a passive machine or tool [24]. A strong sense of social presence usually indicates a more natural and engaging interaction, which can in turn increase trust and empathy towards the agent. Conversely, if the agent is seen as having zero social presence, the user might not engage socially or might not heed social cues from the agent.

934 • **Acceptance** measures how willing users are to interact with the agent and integrate it into
935 their routines [37]. As such, it is an important proxy for the overall quality of the system, as
936 it gauges the possibility of maintaining long-term interactions within the system.

937 One fundamental issue with the measure of these self-reported metrics is the fact that they are
938 traditionally uni-directional, i.e., from the perspective of the human. They are usually collected
939 via questionnaires or interviews before and/or after an interaction, capturing the human’s personal
940 experience, attitudes, and perceptions of the agent or the interaction. Moreover they are intrinsically
941 task-dependent, as they either measure the quality of a specific interaction, or use the result of a given
942 interaction to condition the human to predict the quality of future interactions.

943 Besides self-reported metrics, researchers in human-robot and human-computer interaction often
944 also rely on objective behavioral interaction metrics, i.e., what users do during the interaction. For
945 example,

946 • **Eye gaze** can reflect the engagement of the user in a given interaction: users that frequently
947 make eye contact with the agent, or make prolonged eye contact often report a higher
948 engagement [3, 115, 68]. Moreover, since eye contact is also a sign of social connection,
949 some studies consider mutual gaze episodes (when both the agent and the human look at
950 each other) as a metric of rapport [158], or as a factor influencing the decision-making of
951 the human user [10].

952 • **Task fluency** refers to how coordinated and seamless the joint actions of the human and
953 the agent are in the context of a collaborative tasks [56, 55]. A fluent interaction with high
954 concurrent activity (indicating teamwork) and low idle times often correlates with higher
955 acceptance [5] and trust [101] in the agent.

956 • **Interaction duration** can also be considered an important metric to assess the quality of an
957 interaction, especially in long-term scenarios. The frequency of voluntary interactions with
958 the agent can be considered an indicator of engagement and acceptance [62, 66].

959 Several works have built upon these metrics to develop models that predict the user’s behavior
960 during the interaction. Guo and Yang [50] and Chen et al. [23]) explore how to build data-driven
961 models of the user’s trust for specific collaborative decision-making tasks. Similarly, Lee et al. [82]
962 introduces a computational framework to model the engagement of the human user in a robotic
963 storytelling task. We highlight that agents often do not have access to the internal state of the human
964 and measuring precisely how a user may change as a result of an interaction can be intractable [48].
965 Yet, predictive methods of human interaction metrics could, in principle, still be used as a proxy
966 measurement. However, we carefully point out two limitations of these approaches: (i) they are still
967 heavily task-dependent, limiting their usefulness for long-term, open interactions such as the ones
968 enabled by coexistence; (ii) building static models of the behavior of the user, as currently done, will,
969 once again, lead to the development of generic and stagnant agents, as detailed in Section 2.

970 Rather, we join Kamino et al. [61] (see Appendix D, Social robots, for a discussion of this work) to
971 encourage the AI and interaction communities to shift their emphasis from measuring coexistence
972 to *constructing* coexistence. That is, defining the purpose of the artificial agent and refining its
973 performance is meaningful and reciprocal in and of itself. In other words it is not the task-dependent
974 metrics that ensure coexistence but the construction of them.

975 D Are Current Embodied Agents Already Coexisting?

976 Naturally, one might question whether current embodied agents are already coexisting with humans.
977 In this section, we present examples and discussions on key challenges inhibiting current agents from
978 being coexisting.

979 **Social robots** A prominent example of embodied agents designed for human interaction are social
980 robots [17, 85]. Companies like Jibo and Anki introduced social robots to the market with high
981 expectations, only to face eventual failure [138]. A significant factor contributing to this is the
982 challenge of sustaining long-term interactions by current embodied agents. Without the ability to
983 change through interaction and become situated into their environment, social robots remain ill-suited
984 for prolonged use. They often succumb to the *novelty effect*, where user engagement diminishes over

time as the robot’s initial appeal wears off [118]. The ethnographic field study on social robots by Kamino et al. [61] highlights a notable exception to this trend. Their empirical evidence suggest that meaningful, long-term interactions with embodied artificial agents has less to do with specific design features. Instead meaningful interactions arise from “interconnected moments of situated interaction, related emotional responses, and meaning-making among people as they interact with robots and each other at different levels of organization.” We see this exception as opportunity to design coexisting agents that respond to the broader social and environmental context that they are situated in.

Bias-amplifying interaction Large language models have been widely integrated into the architecture of embodied agents [148, 34]. These models have now been broadly adopted by diverse user groups. While most AI systems influence human behavior, they themselves do not retain user-driven modifications beyond the immediate context window. This lack of adaptability is already problematic, as user-provided knowledge is not incorporated. Worse, studies have shown that interacting with slightly biased AI systems can amplify biases in users, an effect not observed in human-human interactions [47]. These systems not only fail to adapt through interaction, reinforcing a unilateral dynamic, but they also degrade overall system quality by increasing bias in users. As LLMs are increasingly integrated into interactive robots, these issues are likely to persist, if not worsen, through prolonged human-robot interactions.

Please, just turn on the light In industrial settings, robotic failures require expert technicians to debug classifiers, diagnose issues, and retrain models with additional data, such as images captured under varied lighting conditions. Consider now, instead, a robot designed to tidy up homes and offices by identifying, classifying, and sorting objects. Relying on expert interventions, similar to those in industrial settings, is impractical for home-deployed robots. A more viable solution is for robots to make use of humans’ situated knowledge within their environment. Humans understand their space and might recognize how the specific lighting affects object classification. Instead of requiring an expert to retrain the system, a robot could ask for help [64], prompting users to turn on the light and even learning that doing so improves classification performance. By adapting through situated interactions, the robot avoids repeated failures and reduces the need for costly expert intervention and large-scale data collection. This behavior realistically adjusts the human’s expectations on the agent’s capabilities and invites them to accommodate their limitations. Furthermore, we argue that such situated interactions not only improve performance but drive continual evolution and optimization.

E Limitations of Current Learning Algorithms for Coexistence

In this section we argue that, in their current form, standard learning approaches for building embodied artificial agents, designed to overcome the assumption of a static optimization problem, are still insufficient to foster coexistence.

Meta-learning is often proposed as a solution for adaptation to new tasks. A meta-learning agent trains over a distribution of tasks so that it can quickly adapt to unseen task at test time [57]. However, instead of assuming a fixed task (as in an MDP), meta-learning assumes a fixed meta-distribution of tasks, defined by the designer, from which both training and testing tasks are drawn [39]. Consequently, if the agent encounters a fundamentally novel task outside this distribution, it may not be able to adapt, as it was never optimized for tasks beyond the anticipated variations.

Continual learning algorithms, instead, relax the fixed training assumption of the previous methods by exposing agents to a sequence of tasks over time. Often, continual learning algorithms focus on mitigating catastrophic forgetting and preserving performance across task sequences [144]. However, similar to meta-learning, most continual learning setups still operate under relatively controlled forms of novelty: they assume a sequence of tasks or data that, while possibly distinct, follows a predictable format (for instance, new classes from the same data distribution [67]). The focus is on incremental change, not the *unknown unknowns* [83] agents may face in the real-world.

Reinforcement learning (RL) is traditionally formulated through a Markov decision process (MDP), where the state and action spaces, reward and transition functions are defined apriori by the designer [131]. Additionally, the training and testing environment are often assumed to be the same. The goal of the agent is therefore to maximize the total (discounted) reward accumulated while acting on that particular environment. Other extensions, such as partially-observable MDPs to deal with incomplete state information [60], multi-agent MDPs to deal with multiple agents acting in the same

environment [49], domain randomization to deal with environments with different dynamics [25] or observations [153], or robust MDPs to introduce safety guarantees in the behavior of the agent [90], still leave the burden on the researcher to guess the possible perturbations the agent might face in advance. This is often done by formalizing uncertainty in *quantitative terms*: assuming any environment change can be modeled as some stochastic variation in rewards, transition probabilities, or observation noise. However, these approaches are still unable to deal with structural novelty that might lie outside the prior distribution of possible phenomena defined by the designer. Recently, continual reinforcement learning methods, while still severely underexplored [113], have shown some promise in allowing agents to carry out search processes indefinitely [1], producing behavior in response to all past experience [16]. We believe that these approaches could also be used in the future, complementing evolutionary techniques, to allow agents to explore convergent and divergent behaviors.

For an extended argument on the limitations of these methods for coexistence, we refer the reader to Lehman et al. [83].

F Potential Coexisting Technology Today

In this section, we highlight several examples of technology with properties that foster coexistence.

Mutable morphology and locomotion Figure 4 shows how the morphology of an agent can be changed to recover from damage and to re-learn how to walk [71]. The agent learns how to walk through periodically inflating and deflating its individual cells, exploiting its own physical shape. Although this does not involve a human user, it demonstrates the value of mutable morphologies. For example, we see great potential in mutable morphology to express various mannerism through different gaits, especially in the context of Vergunst and Ingold’s work on the contextual nature of walking [142]. Thus culminating in rich, heterogeneous populations of artificial agents at scale. Figure 5 shows Yamaha’s “Motorid”, a shape changing, self-balancing motorcycle [52]. It has a twisting chassis and autonomous driving abilities that influence how riding the motorbike feels in real time. This dramatically changes motorcycling from its culture to its engineering principles. Whilst not a child of the RtD method, but rather a concept bike, it balances divergent and convergent themes. Blurring the definition of what is a bike and an autonomous agent.

Mutable perceptive fields in human-drone interaction Figure 6 shows a system that allows humans to shape how a drone senses its environment. So whilst the human understood how to shape the perceptive field of the drone, the behavioral outcomes of their actions were not. This enabled rich explorations with the drone which lead to diverse and meaningful interactions [78].

Mutable interaction interface Figure 7 shows Bewley and Boer’s [11] “Blo-Nut”, a silicone doughnut that affords the user a blank slate to interact with. The object inflates and deflates and can be programmed to music. Its non-humanoid shape enables interactions to the human in ambiguous ways, which Sandry [122] argues, is an opportunity to build effective communication between humans and artificial agents.

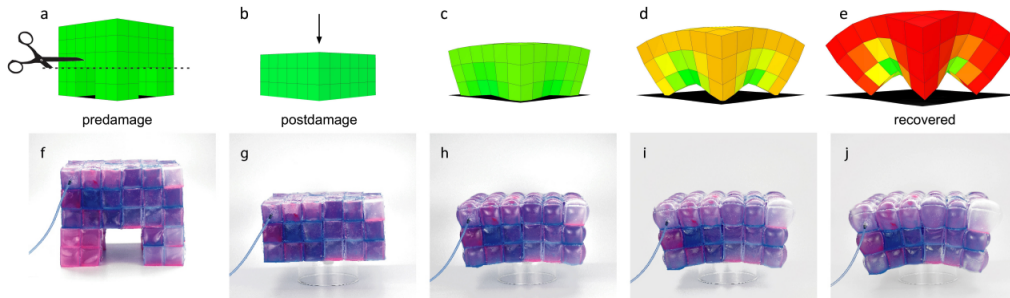


Figure 4: Self recovering locomoting voxels [71]: by virtue of an evolutionary algorithm, the agent is relearning how to walk by changing the inflation patterns of its individual cells. Each change to the physical body is likened to a divergent search for a new and unique locomotion gait. Whilst each improvement in performance in this new body is a convergent search to optimize. Therefore, the agent is able to explore and exploit its physical environment by simply existing in it.



Figure 5: Yamaha's "MOTOROiD" is a shape changing, self-balancing motorcycle [52]. Its unique twisting chassis is able to affect the ride feel in real time as well as drive autonomously.



Figure 6: “How to Train Your Drone” [78]: depicted here in orange, clear and blue are the sensory fields of the drones. By interacting with the drone, its sensory field can be changed with human intention. However the consequences of such changes are not always predictable. This work demonstrates the potential of interacting with the sensing and acting capabilities of mutable agents.



Figure 7: “Blo-Nut” is a silicone doughnut that affords the user a blank slate to interact with [11]. The object inflates and deflates and can be programmed to music.

1075 **NeurIPS Paper Checklist**

1076 **1. Claims**

1077 Question: Do the main claims made in the abstract and introduction accurately reflect the
1078 paper’s contributions and scope?

1079 Answer: [\[Yes\]](#)

1080 Justification: In this paper we introduce, define and argue in favor of the concept of
1081 coexistence for the design of interactive embodied artificial agents. In Section 2 we discuss
1082 the limitations of current embodied artificial agents. In Section 3 we formally define
1083 coexistence and its properties. In Section 4 we look into evolutionary biology and design
1084 theory for inspiration on the mechanisms that can foster coexistence. Finally, in Section 5
1085 we present key research directions for the AI community to develop coexisting agents, and
1086 discuss the ethical implications of such choice.

1087 Guidelines:

- 1088 • The answer NA means that the abstract and introduction do not include the claims
1089 made in the paper.
- 1090 • The abstract and/or introduction should clearly state the claims made, including the
1091 contributions made in the paper and important assumptions and limitations. A No or
1092 NA answer to this question will not be perceived well by the reviewers.
- 1093 • The claims made should match theoretical and experimental results, and reflect how
1094 much the results can be expected to generalize to other settings.
- 1095 • It is fine to include aspirational goals as motivation as long as it is clear that these goals
1096 are not attained by the paper.

1097 **2. Limitations**

1098 Question: Does the paper discuss the limitations of the work performed by the authors?

1099 Answer: [\[Yes\]](#)

1100 Justification: In Section 5.1 we discuss the ethical implications of the development of
1101 coexisting agents. Furthermore, in Section 6 we present alternative viewpoints to coexistence
1102 for the design goals of embodied agents.

1103 Guidelines:

- 1104 • The answer NA means that the paper has no limitation while the answer No means that
1105 the paper has limitations, but those are not discussed in the paper.
- 1106 • The authors are encouraged to create a separate "Limitations" section in their paper.
- 1107 • The paper should point out any strong assumptions and how robust the results are to
1108 violations of these assumptions (e.g., independence assumptions, noiseless settings,
1109 model well-specification, asymptotic approximations only holding locally). The authors
1110 should reflect on how these assumptions might be violated in practice and what the
1111 implications would be.
- 1112 • The authors should reflect on the scope of the claims made, e.g., if the approach was
1113 only tested on a few datasets or with a few runs. In general, empirical results often
1114 depend on implicit assumptions, which should be articulated.
- 1115 • The authors should reflect on the factors that influence the performance of the approach.
1116 For example, a facial recognition algorithm may perform poorly when image resolution
1117 is low or images are taken in low lighting. Or a speech-to-text system might not be
1118 used reliably to provide closed captions for online lectures because it fails to handle
1119 technical jargon.
- 1120 • The authors should discuss the computational efficiency of the proposed algorithms
1121 and how they scale with dataset size.
- 1122 • If applicable, the authors should discuss possible limitations of their approach to
1123 address problems of privacy and fairness.
- 1124 • While the authors might fear that complete honesty about limitations might be used by
1125 reviewers as grounds for rejection, a worse outcome might be that reviewers discover
1126 limitations that aren’t acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: While we do not have theoretical proofs, we do present in full our formalism of coexistence in Section 3. Moreover, we present additional considerations, such as consequences, assumptions and limitations of our formalism in Section B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: We do not present experimental results in this paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

1181 (d) We recognize that reproducibility may be tricky in some cases, in which case
1182 authors are welcome to describe the particular way they provide for reproducibility.
1183 In the case of closed-source models, it may be that access to the model is limited in
1184 some way (e.g., to registered users), but it should be possible for other researchers
1185 to have some path to reproducing or verifying the results.

1186 5. Open access to data and code

1187 Question: Does the paper provide open access to the data and code, with sufficient instruc-
1188 tions to faithfully reproduce the main experimental results, as described in supplemental
1189 material?

1190 Answer: [NA]

1191 Justification: We do not present experimental results, neither code nor data.

1192 Guidelines:

- 1193 • The answer NA means that paper does not include experiments requiring code.
- 1194 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/](https://nips.cc/public/guides/CodeSubmissionPolicy)
1195 [public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1196 • While we encourage the release of code and data, we understand that this might not be
1197 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
1198 including code, unless this is central to the contribution (e.g., for a new open-source
1199 benchmark).
- 1200 • The instructions should contain the exact command and environment needed to run to
1201 reproduce the results. See the NeurIPS code and data submission guidelines ([https://](https://nips.cc/public/guides/CodeSubmissionPolicy)
1202 nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- 1203 • The authors should provide instructions on data access and preparation, including how
1204 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 1205 • The authors should provide scripts to reproduce all experimental results for the new
1206 proposed method and baselines. If only a subset of experiments are reproducible, they
1207 should state which ones are omitted from the script and why.
- 1208 • At submission time, to preserve anonymity, the authors should release anonymized
1209 versions (if applicable).
- 1210 • Providing as much information as possible in supplemental material (appended to the
1211 paper) is recommended, but including URLs to data and code is permitted.

1212 6. Experimental setting/details

1213 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
1214 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
1215 results?

1216 Answer: [NA]

1217 Justification: We do not present experimental results.

1218 Guidelines:

- 1219 • The answer NA means that the paper does not include experiments.
- 1220 • The experimental setting should be presented in the core of the paper to a level of detail
1221 that is necessary to appreciate the results and make sense of them.
- 1222 • The full details can be provided either with the code, in appendix, or as supplemental
1223 material.

1224 7. Experiment statistical significance

1225 Question: Does the paper report error bars suitably and correctly defined or other appropriate
1226 information about the statistical significance of the experiments?

1227 Answer: [NA]

1228 Justification: We do not present experimental results.

1229 Guidelines:

- 1230 • The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: We do not have experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and our work conforms fully to it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In Section 5.1 we discuss the ethical implications of the development of coexisting agents.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We do not use any existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not provide any novel assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not present any research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not present any research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- 1385 • Depending on the country in which research is conducted, IRB approval (or equivalent)
1386 may be required for any human subjects research. If you obtained IRB approval, you
1387 should clearly state this in the paper.
- 1388 • We recognize that the procedures for this may vary significantly between institutions
1389 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1390 guidelines for their institution.
- 1391 • For initial submissions, do not include any information that would break anonymity (if
1392 applicable), such as the institution conducting the review.

1393 16. **Declaration of LLM usage**

1394 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1395 non-standard component of the core methods in this research? Note that if the LLM is used
1396 only for writing, editing, or formatting purposes and does not impact the core methodology,
1397 scientific rigorousness, or originality of the research, declaration is not required.

1398 Answer: [NA]

1399 Justification: We do not use LLMs in this paper as a core method of our research.

1400 Guidelines:

- 1401 • The answer NA means that the core method development in this research does not
1402 involve LLMs as any important, original, or non-standard components.
- 1403 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1404 for what should or should not be described.