5

Synthetic Collaborative Systems

Present and Future

Ismael T. Freire and Paul F. M. J. Verschure

Abstract

Collaboration in synthetic systems can inform our understanding of collaboration as a natural phenomenon. A general overview of how collaboration has been studied in the human evolutionary behavioral sciences is presented, and it is argued that further synthesis is needed across the different levels of explanation. At present, two main issues prevent such a synthesis: (a) the current gap between proximate accounts of collaboration from the cognitive sciences and the ultimate levels of explanation from the evolutionary sciences, and (b) methodological limitations which inhibit joint study of collaboration at different levels of description. Synthetic collaborative systems (e.g., robotics and artificial intelligence) can help to address these issues. A unified research program centered on *synthetic collaboration* provides a way to expand understanding of human and animal collaboration, to design and study human–machine collaboration, as well as to investigate purely synthetic forms of collaboration between intelligent machines. Here, current research is reviewed that employs synthetic methodologies across different fields, the implications of developing synthetic collaborative agents are discussed, and an approach is proposed to study both natural and synthetic collaboration, under the name of *collaborative cybernetics*.

> Science requires an almost complete openness to all ideas. On the other hand, it requires the most rigorous and uncompromising skepticism. —Carl Sagan (1995)

Introduction: The Dusk Colony

Imagine the following: The first human colony on another planet, named after its main private sponsor, Dusk, was constructed over a seven-year period on our neighboring planet, Mars. The green tips of the pyramid-shaped buildings can be clearly seen from the planet's surface, proudly rising from the ground,

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

like a set of icebergs floating in a sea of red sand and rocks. Underground, a complex system of interconnected caves has been carved out to protect the earthling society from the hostile environment above, with its high levels of ionizing radiation and extreme range in surface temperature. Autonomous 3D printed and 3D printing factories process the extracted materials from a nearby mining station to produce concrete from sand and rock for industrial-scale terraforming. The main engine of the spaceship that brought the settlers to Mars has been repurposed as the central power plant to provide energy to core infrastructures and to kick-start the production of solar cells. Through these efforts, the surface of the red planet is slowly turning green. Greenhouses stretch for miles, filled with a combination of terrestrial flora in hydroponic plantations enhanced with artificial photosynthesis systems that sequester carbon dioxide from the atmosphere and produce oxygen. Additionally, a novel prototype plant transforms briny water from the surface into water and oxygen to feed the greenhouses. A large water extraction system takes advantage of a nearby deposit of frozen water below the surface. Experiments with genetically precision-engineered organisms have added mobility and nervous systems to plants, allowing them to adjust to extreme conditions. Thanks to these innovations and thousands of human-years of work, the first extraterrestrial city is emerging to provide shelter to the exponentially growing population. Through these efforts, Mars is becoming a springboard for the further population of the solar system and beyond, effectively making humanity a space-bearing, multi-planetary species. As in all cities, Dusk provides shelter, protection from radiation and extreme temperatures, energy, oxygen, water, and food to a self-sustaining and rapidly growing population. The colony has everything, except for one thing: humans.

The Dusk colony has been built by and is entirely comprised of robots—robots capable of maintaining themselves under extreme conditions, auto-organizing, efficiently distributing work among themselves to accomplish the larger goals of building and maintaining the Mars colony and preparing colonization missions to the next extraterrestrial footholds for the descendants of human civilization: the moons of Saturn and Jupiter, Titan and Europa, respectively. Dependence on centralized terrestrial control, be it human or artificial intelligence (AI), is not an option for various reasons: the radio transmission delays between so-called Spacecraft Event Time and Earth Received Time (i.e., between four and 24 minutes), the lack of operator situational awareness in teleoperation, and the chaos that would ensue when thousands of human and synthetic teleoperators try to coordinate their intentions and actions. Autonomous operation at both the individual and collective level is thus a necessary condition for Dusk. Indeed, being so far away from human oversight, in the deep vastness of space, no one can fix a *blue screen of death*. Therefore, the robots that engage in space exploration need to be both robust (to withstand the harsh conditions of the Mars environment) and autonomous (to control their individual and collective actions). Above all, if they are to succeed in realizing

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

their common goal of colonizing Mars and beyond, they must achieve unprecedented levels of *collaboration*.

The colonization of space may lay within the *adjacent possible*, but what is required to make it an actuality? Before addressing the future of space exploration and colonization, humans must first channel the collaborative efforts of our entire species and its technologies to solve the most pressing issue of our time, or any time: the human-driven ecological collapse of Earth. To realize synthetic collaboration necessary for space colonization and to answer the acute threats caused by human's tendency toward self-destruction, we must first understand the nature of collaboration as displayed by biological systems, such as social insects, mammals, and especially *Homo sapiens*. To this end, we propose that our understanding of collaboration can be greatly enhanced by constructing synthetic collaborative systems, such as those of the imaginary Mars settlement Dusk. Indeed, synthesizing and understanding collaboration are codependent and interlinked endeavors: the development of synthetic collaborative agents will both shape and amplify human collaboration and play a key role in understanding the nature of natural collaboration and its underlying principles, following the *factum et verum* principle of Giambattista Vico: We can understand that which we can build (Verschure 2012).

The Many Faces of Collaboration

This Ernst Strüngmann Forum was driven by basic questions: What is collaboration? What is it good for? Why do we collaborate? As a starting point, collaboration was understood as "cooperation between agents toward mutually constructed goals" (Roepstorff and Verschure, pers. comm.). Accordingly, we must understand the constituent concepts: cooperation, agents, and goals.

More than 150 years after *The Origin of Species* (Darwin 1909), unraveling the mysteries of cooperation remains one of the most challenging aspects of animal behavior. Indeed, Darwin considered cooperation as a puzzle to his theory of natural selection. When natural selection favors the evolution of behaviors that improve an individual's fitness or reproductive success, cooperation generally implies that the recipient of the cooperative action would improve its fitness at the cost of the donor. This relationship contradicts the general reductionist premise behind natural selection (Apicella and Silk 2019).

Natural cooperation can be studied at different levels of description: from the biophysical and neuronal (as happens in bonding through the hormone oxytocin and the switching of circuits in the hypothalamus), to the emotional recognition of social cues, the cognitive processes of theory of mind (ToM) and social decision making, up to the population dynamics of collective behavior and the shaping forces of the physical, social, and cultural environment. To address these different aspects, cooperation is studied in multiple disciplines: in the natural and social sciences, in engineering, as well as in the

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

field of artificial intelligence. This multidisciplinary, partially reductionist, and fragmented decomposition of the phenomenon automatically generates a new challenge; namely, how can insights from these different fields be integrated into one comprehensive perspective? We hold the realization of synthetic collaboration can be a fruitful method to push forward this epistemic integration (Prescott et al. 2018).

The View from Biology

Any biological phenomenon must be viewed through the lens of evolution, as Dobzhansky (2013) famously recommended. Yet, what is cooperation exactly? Following Mayr (1961) and Tinbergen (1963), the quest to explain any behavior can be divided into (a) the proximate causes of its ontogeny and the immediate mechanical specification of a trait and (b) the ultimate causes of the phylogeny and function of such a trait. The latter provides the evolutionary explanation for the former in a unidirectional causal pathway. This perspective is a distinguishing feature of the so-called Modern Synthesis in evolutionary biology, which integrated natural selection, population genetics, and Mendelian inheritance. Effectively, this created a tension with the Aristotelian material, formal, and final causes, which had fallen in disrepute through the rise of physics in the nineteenth century and the associated reductionist dream of unified science driven by unidirectional feedforward causality. Currently, Modern Synthesis is facing challenges that dislodge it further from these reductionist roots toward a multiscale systems perspective based on bidirectional and recurrent causal interaction. These challenges result from new insights in the dynamics of evolutionary processes, including

- the tight coupling between development and evolvability in epigenetics (or evo-devo), where phenotypic change is seen to result from environmental perturbations engaging with developmental programs rather than feedforward genetic mutations,
- inclusive inheritance through parental shaping of developmental resources, from metabolism and behavior to the selection of hosts and environments, and
- niche construction, where organisms change their environment to the extent that it, in turn, affects selection, thus leading to ecological inheritance (for a review, see Laland et al. 2015).

As a result, an "extended evolutionary synthesis" has emerged to question the unidirectional causal relationship between ultimate and proximate causes solely driven by selection upon gene variants. We argue that cooperation and collaboration add yet another and possibly more profound challenge to the Modern Synthesis by not only emphasizing the multiscale dynamics of the organism and its niche but adding the further amplification derived from the dynamics of

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

collectives of collaborating agents with access to internal and external symbolic memory and transfer. Collaboration establishes new causal pathways for niche construction, adaptation, and selection (see also Chapters 12 and 16, this volume). Here, we elaborate on the specific proposal that due to the critical role of multiscale feedback loops in processes of collaboration, new methods must be devised to understand, study, and explain collaboration that are complementary to current quantitative and qualitative approaches. The construction of synthetic collaborative systems provides an effective approach to bridge the gaps in our understanding because they allow, potentially, for a more precise experimental control of the many feedback loops and spatiotemporal scales involved. The construction of the Dusk colony would thus not only be an engineering feat, but also a step toward advancing a fundamental understanding of ourselves as a profoundly collaborative species.

Several evolutionary mechanisms have been proposed to explain the emergence of cooperation (Nowak 2006). Kin-based altruism explains cooperation between genetically related members of a species through the introduction of the construct of inclusive fitness, where fitness is understood as a measure of reproductive success (Hamilton 1964). In this view, the cost of assisting kin in terms of foregone reproduction can be offset and compensated if it is beneficial to genetically related members. This means that genetic relatedness provides a necessary condition for this form of cooperation. Direct reciprocity, or reciprocal altruism, extends the reach of cooperation beyond kin (Axelrod and Hamilton 1981; Trivers 1971). In a context where repeated interactions with the same individual might occur, reinforcement (including the possibility for punishment) allows cooperative strategies to be maintained over time even with nonrelatives. Such reinforcement-based mechanisms follow the classic Law of Effect advanced by Thorndike (1927): behavior is shaped through its outcomes or reinforcement, which has been shown to hold only for a very limited set of adaptive behavior. Indirect reciprocity, with the addition of virtual reinforcers such as reputation and signaling, allowed for these reciprocal mechanisms to extend the reach of cooperation beyond that of directly known individuals of a group and explain its potential scaling. This has been well illustrated through the analysis of small primate communities, such as bonobos and chimpanzees (de Waal 1989; de Waal and Lodge Jr. 1973).

Unfortunately, despite its success in explaining these key forms of cooperation, an evolutionary gene-centric view does not seem sufficient to account for the breadth of human cooperation. The emergence of social norms and institutions in human societies poses a challenge to this paradigm, as neither kin-based altruism nor direct reciprocity can alone explain such unprecedented levels of cooperation (Boyd and Richerson 1985). By focusing on genes and natural selection, the modern evolutionary synthesis left other types of evolutionary processes out of the equation.

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

Research from cultural evolution studies on human cooperation and behavioral diversity further contributes to the extended evolutionary synthesis, an example of multigenerational behavioral feedback and niche construction. The Interdependence Hypothesis states that the unique forms of human cooperation and cognition arose in two steps (Tomasello et al. 2012). In a first step, ecological forces pressured early humans to become collaborative foragers. By virtue of this constraint, we became interdependent: it was in an individual's best interest to care about the well-being of dependent others, since it was a necessary condition for achieving mutual goals serving survival (e.g., foraging, hunting, or defending the group). Achieving such collaborative interactions required the refinement of new cognitive capabilities, such as joint intentionality and mind-reading, to facilitate the alignment of individuals in a bottom-up fashion to form a collective. In addition, a new type of top-down factor came into play that enhanced the coherence in collective processes in the shape of a second-person normativity or joint morality derived from the cooperative process itself (Tomasello and Vaish 2013). In a second step, groups of hominids had to compete with other groups and species for space and resources. This between-group competition extended the functional interdependence from small coalitions to large collectives, which required a new form of collective intentionality or group-mindedness. The hypothesis is that this transition from agent-dependent cooperation to group-level and agent-neutral collaboration formed the foundations of social conventions, norms, and institutions. It is believed that this transition played a critical role in the ability of *H. sapiens* to overcome other hominids, in particular *H. erectus*. Over time, competition between larger groups of hominids that held different norms and institutions led to a cultural evolutionary selection process in which the groups with more cooperative norms managed to outcompete their rivals. In brief, this between-group competition led to within-group cooperation in large-scale societies. From this perspective, cultural change is itself regarded as an evolutionary process, in which cultural traits are formed and spread according to their utility, attractiveness, and compatibility with existing traits and diversify through a cumulative process of invention, selection, elaboration, and refinement (Brown et al. 2011). In this process, we can observe a transition from mentalizing at the level of agents and their interactions, to collectives built by agents, to abstract institutions. Hence, at the heart of this complex form of collaboration stands the shaping of intentions, their representation by agents, and their externalization and formalization. This implies that collaboration plays out in the models that agents and collectives maintain of their physical and social environment, their tasks, and each other, rather than just in the physicality of existence and interaction itself. We refer to this ability as the *capability for virtualization*.

Despite its progress toward further integration, every synthesis leaves something behind. The extended evolutionary synthesis is no exception: the broader cognitive sciences, and the study of the proximate cognitive mechanisms that

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

give rise to cumulative cultural evolution have been omitted. Indeed, the characteristic ratcheting process of cumulative cultural evolution relies critically on a set of cognitive mechanisms including memory, consciousness, language, metacognition, ToM, social perception, selective social learning, teaching, and norm psychology, to name but a few (Birch and Heyes 2021). Further advances, therefore, should focus on taking such cognitive processes into account. In this new integrative view, evolution by natural selection could be seen as a more general process of variation, selection, and heredity unfolding in multiple dimensions: genetic, epigenetic, behavioral, and symbolic (Deacon 2011; Jablonka and Lamb 2006; Thierry 2007). In the study of the feedback loops between these four dimensions that link the proximate and the ultimate causes and consequences of collaboration, the use of synthetic collaborative agents will be extremely important because they will allow for the control of time and a third-person analysis of first-person states (Verschure 2016b). Hence, we observe a paradoxical transition in the biological perspective on evolution and collaboration, where the initial objective reductionist ambition heralding a new science of life has led us down a path where the subjective states of the agents that form collaborative groups must become central ingredients of the explanation. In other words, by moving the explanation of evolution from the directly observable physical world to the only indirectly accessible social world of collaboration, the dominant paradigm of evolution included in its extended form has become incomplete (Deacon 2011). Precisely because of the inclusion of "the other" and the necessary virtualizations underlying collaboration, we propose that synthetic methods will be key as they allow third-person access to first-person states of thinking and experiencing synthetic agents (Verschure 2016b).

The Relevance of a Synthetic Approach

To fill the epistemic gap in our understanding of collaboration, further steps need to be taken. We propose that the study of synthetic collaboration can help bridge the gap between the multiple levels of organization that underpin collaboration. By building artificial agents that can collaborate with us as well as among themselves, based on biological principles of control and communication, we will deepen our knowledge of collaboration and lay the groundwork for a new class of collaborative technologies that may help us achieve the colonization of Dusk on Mars and beyond.

The philosophical roots of the synthetic approach proposed here date back to the *factum et verum* principle of Giambattista Vico (1999). For Vico, creation was a form of understanding. Thus, building synthetic collaborative agents can bring several benefits as well as create new challenges (Lallee et al. 2015; Verschure 2016b). First, it will help us build better models of human and animal collaboration. Building something from scratch will help us identify

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

the missing gaps in knowledge, focus on the underlying principles, and impose real-world constraints upon our developing theories, opening them up for empirical proof-of-concept validation. Second, it gives us unique access to the inner workings of a system, such as the memory states that provide the substrate for virtualization as deployed in mental time travel and mentalizing. This will permit us to investigate a behaving adaptive system from a privileged vantage point, as we will have direct access to the activity of its cognitive system and its full behavioral and experiential history. It will also enable the researcher to identify direct connections between environment, embodiment, cognition, and behavior. Third, the development of synthetic agents opens a new branch in the study of collaboration itself: that of artificial systems in machine–machine, machine–human, and machine–animal forms. This will allow us to explore new scenarios beyond the constraints of biological systems, such as memory capacity, communication bandwidth, or perceptual processing. By focusing on fully synthetic systems, we can move beyond biological plausibility in terms of physical and computational capabilities. Fourth, traditional empirical noncomputational methods struggle to study collaboration simultaneously at different levels of description. Studies that focus on tracking brain activity during social interaction in controlled environments (e.g., Hamilton 2021; Li et al. 2021; Yang et al. 2020) are usually conducted between dyads or very small groups of participants to assure controllability. Extending the study of such proximate levels to larger populations, while balancing the trade-off between experimental control and ecological validity, is feasible but not a trivial task given our current technology.

The Paths Created by Synthetic Collaboration

In addition to the arguments above, a synthetic approach to study collaboration offers additional advantages. It would, for instance, create several new concrete research avenues and permit the theoretical aspects of collaboration to be more effectively studied, as discussed below.

Embodiment, Affordances, and the Morphospace of Collaboration

Within the cognitive sciences, embodied cognition theories emphasize the active role of the physical body in shaping cognition (Clark 1999; Wilson 2002; Shapiro 2010). From this perspective, an agent's cognition is (at least partially) shaped and determined by its body as well as its interactions with the environment. This consideration can be applied to any type of cognitive agent, biological or synthetic. This means that the properties of an agent's physical structure determine what information it can perceive and which actions it can perform (i.e., its perceptual and motor systems). Body specifications also determine what is required for its proper maintenance. These physical and physiological constraints, in turn, shape the agent's motivational and cognitive systems:

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

how it thinks as well as what it can do. Borrowing from embodied cognition theories, the notion of affordances in human–robot interaction studies refers to action possibilities that the environment provides to an agent (Moratz and Tenbrink 2008). In short, what an agent can do at any given moment is fully determined by the limits and possibilities of the agent's body and the ecological constraints of the environment in which it's embedded. For instance, an agent endowed with a tool that extends its reach can grab objects further away than it would normally be able to do.

Combining these two notions, collaboration can be thought of as a way for each agent to extend its affordance space beyond the limitations of its current capacities. In other words, collaboration allows agents to achieve goals beyond the reach of a single individual. This holds true even for collaboration between organisms of the same species who share similar affordances due to their comparable body configuration. This point becomes even more relevant when we think of how collaboration between agents with different body compositions and configurations could capitalize on the particular skills and capacities of each individual agent. Human–robot collaboration falls within this category. Examples of this can be seen in search and rescue missions in which humans and drones collaborate and benefit from each other's capabilities. Beyond what agents with different bodies can achieve by working together, there are other possible collaboration regimes that entail cognitive collaboration. Collaboration between agents with different knowledge sources can render benefits greater than the sum of their individual capacities. In future research on synthetic collaboration, this needs to be explored, given its implications for the design and development of multi-robot activities (e.g., the Dusk colony); that is, situations where different types of synthetic agents endowed with diverse morphologies and skills need to interact together.

Indeed, purely novel synthetic forms of collaboration can be developed due to the interaction of agents that did not previously exist in the natural world. In biology, the concept of a morphospace defines a graphical representation of all the morphologies an organism could or does have; each point represents an individual shape (Arsiwalla et al. 2017; Ollé-Vila et al. 2016). The development of novel forms of embodied artificial intelligence, made of completely different materials to biological organisms, may give rise to agents that can perform physical and cognitive tasks in unique ways. The development of synthetic agents—from robots that can carry heavy loads to AIs that can compute complex algorithms in fractions of a second—opens the possibility of extending the morphospace of collaboration.

The Creation of a Synthetic Umwelt

The nineteenth-century philosopher Jakob Von Uexküll coined the term *Umwelt* to refer to the perceptual world in which an agent exists and acts as a subject

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

(Von Uexküll 1992). It also refers to the semiotic world, as it includes the meaningful aspects of the world for any particular agent. In contrast to the notion of *Umwelt*, the term *Umgebung* describes what the *Umwelt* is as viewed by an observer. If the *Umwelt* is the agent's first-person, bodily constrained, subjective experience of the world, the *Umgebung* is the third-person description of such an experience. In the study of biological systems, there has always been an insurmountable barrier between these two concepts. By developing artificial agents, endowed with a specific set of sensors and actuators, we would de facto be engineering an *Umwelt*. However, unlike their biological counterparts, we would have complete access to the integrated first-person perspective of the synthetic agents. In other words, we would dismantle an epistemological barrier that has not yet been breached, thus creating a possibility to study the relationship between an agent's body, its environment, and the first-person experience of both.

In summary, the materials and body plan of a synthetic agent determine its needs, capacities, and first-person experience of the world. This, in turn, affects the space for possible collaborative acts that the agent will be willing and capable of performing. Fully understanding the relationship between the agent's body, cognition, and *Umwelt* is of paramount importance for designing the type of collaborative agents with which we want to share the world.

Multilevel Collaboration and Open-Endedness

The use of synthetic methods can enable a proper multilevel study of collaboration. If we had direct access to real-time information from cognitive, behavioral, population, and ecological levels, our analysis could go beyond simple correlations and begin to focus on causation between levels. Moving in this direction, novel methodologies have been developed to study downward and bottom-up causation between different levels of description (Hoel 2018; Klein and Hoel 2020; Rosas et al. 2020), thus allowing us to determine more accurately what level of description is more informative for the study of a target phenomenon and to better classify the emergent phenomena (Varley and Hoel 2022).

This opens the possibility of studying how certain environmental dynamics causally affect the behavior and cognition of different species. For instance, we can investigate under which ecological conditions certain behaviors are expressed and gain a deeper understanding of what drives behavioral convergence across species (Barsbai et al. 2021). In the same direction, future research could study how different cognitive mechanisms may give rise to similar collaborative behaviors (Raihani 2021). It would be extremely useful to understand how different cognitive paths can reach the same behavioral destination if we aim to introduce synthetic collaborative agents into our daily lives.

A third aspect of the multilevel study of collaboration relates to the study of open-ended evolution and adaptation. The bidirectional feedback loops

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

between different levels of description have been proposed to be one of the driving forces of open-ended evolution in biological systems (Birch and Heyes 2021; Dunbar 2003; Muthukrishna et al. 2018). In-depth studies of this process would be possible using synthetic agents. Moreover, such studies could help us to understand whether the same process takes place in the development of synthetic agents per se, so that we could extract more general and underlying computational principles. Studying the generative process that gives rise to the never-ending novelty of forms, intelligence, and behaviors seen in nature might also open the door to understanding the open-ended nature of human–machine collaboration.

Synthetic Collaboration: The Present

A Fragmented and Diverse Landscape

Extended evolutionary synthesis is advancing rapidly, and new results driving the field forward hold the potential of establishing a true Kuhnian paradigm shift (Kuhn 1970). In contrast, the study and development of synthetic collaborative agents is more fragmented, simplified, and devoid of a unifying research framework.

Agent-based modeling approaches focus on studying population-level dynamics through simulations of large numbers of relatively simple agents. Rooted in the artificial life movement from the late 1980s, this type of modeling work centers on how population-level effects emerge or self-organize from interactions between simple agents (Baronchelli 2018). Agent-based modeling techniques have been widely used in

- traditional (Axelrod 2006) and evolutionary (Smith and Price 1973) game theory to study cooperative and competitive dynamics (Adami et al. 2016; Kaviari et al. 2019),
- cultural evolution studies (Richerson and Boyd 2004) to explain the formation of social norms and conventions (Migliano and Vinicius 2022; Richerson and Henrich 2009) as well as selective social learning (Lewis and Laland 2012; Migliano and Vinicius 2022; Richerson and Henrich 2009; Thompson et al. 2022),
- behavioral ecology to study collective behavior overall (Couzin et al. 2002; DeAngelis and Diaz 2019), and
- linguistics to study the emergence of communication (Steels 2001, 2016; Tseng and Son Nguyen 2020).

Traditionally, the type of agents used deploy minimal cognitive capabilities, as they follow simple rules for behavior and learning or direct heuristics. For instance, agent-based models of cultural evolution are used to study biases

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

in social learning and cultural transmission, such as similarity bias (Saunders 2022) or conforming to majority bias (Youngblood 2019).

In contrast, research on social cognition in artificial agents focuses on the complexities of local dyadic interactions, using more elaborate models to represent the agent's social cognition. This type of modeling centers on the unraveling of the most proximate mechanistic causes of social behavior. Notably, such models are developed in computational neuroscience to study the neural and computational basis of social cognition (Chang et al. 2021; Cushman and Gershman 2019) and decision making (Hutcherson et al. 2015; Olsson et al. 2020) as well as in social robotics to study and develop socially competent artificial agents and its effects on its interaction with humans (Hiatt et al. 2017; Johal et al. 2015; Lallee et al. 2015; Sarathy et al. 2016). These models prioritize biological plausibility over population-level effects and which computational mechanisms, operating under such constraints, can give rise to social cognitive capacities such as social learning, ToM, norm psychology, and even moral decision making (Freire et al. 2020b).

These two approaches show a stark contrast in the use of artificial agents for studying the cognitive mechanisms behind multi-agent and dyadic collaboration: one uses minimal cognitive agents in big numbers, whereas the other uses complex cognitive models in small numbers (Hawkins et al. 2019a). This roughly reflects the transition from the Modern Synthesis to the extended evolutionary synthesis described earlier. Both views come from a long and successful tradition of studies and focus on the level of description suitable for the phenomena being studied. However, a truly synthetic approach that integrates both levels is still missing. Such a multilevel approach is needed to study top-down and bottom-up interactions among cognitive mechanisms that underpin collaboration, local dyadic interactions, and population-level effects. Within the field of artificial intelligence, emerging research on cooperative AI is beginning to tackle this problem (Dafoe et al. 2021) yet with strong roots in traditional agent-based models.

Collaboration in Artificial Intelligence

The emerging multidisciplinary field of cooperative AI aims to develop algorithms that can display cooperation with an emphasis on contemporary AI methods, such as deep learning (Dafoe et al. 2021). This includes creating artificial agents capable of handling cooperative situations, building tools to foster cooperation in populations of agents, and otherwise conducting AI research into problems of cooperation and social AI. The field integrates several research lines and topics within AI captured in the capabilities of social cognition, communication, moral action, and decision making, and building on multi-agent systems, classical game theory, and machine ToM.

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

Multi-Agent Reinforcement Learning

Advanced forms of cooperation require experience-dependent adaptation of behavior and, as a result, popular methods of machine learning are applied to this area. Although the line of research on multi-agent reinforcement learning (RL) is relatively young, its research output is rapidly growing as, in the study of the emergence of conflict and cooperation in agent populations (for extensive reviews, see Busoniu et al. 2008; Gronauer and Diepold 2022; OroojlooyJadid and Hajinezhad 2019). Recently, benchmarks inspired by game theory are becoming standard in the multi-agent RL literature (Freire et al. 2020a; Leibo et al. 2017; Lerer and Peysakhovich 2017; Rabinowitz et al. 2018). Building on this trend, some researchers are translocating the type of conflicts represented in classic game-theoretic tasks (e.g., the iterated prisoners dilemma) into more ecologically valid versions (Lillicrap et al. 2019). One particularly active line of research focuses on extending the Deep Q-Learning network architecture, proposed in Mnih et al. (2015), into the social domain (Leibo et al. 2017; Lerer and Peysakhovich 2017; Perolat et al. 2017). This architecture combines an RL algorithm that labels states of the (social) world in terms of the most rewarding actions with which they are associated. These states are abstract features extracted from raw image pixels by a deep convolutional neural network. Given the decisive role of virtualization in collaboration, such approaches will face limitations. They will, however, be able to delineate an upper bound of cooperation driven purely by surface features that can be picked up by visual sensors, such as position in space, proximity, and posture. For instance, agents that build on this model-free approach—where the agent learns by direct matching of sensory states to actions without resorting to an internal model—are already capable of learning how to play a two-player video game, such as Pong, from raw pixel data and have achieved human-level performance (Mnih et al. 2015) both in cooperative and competitive modes (Tampuu et al. 2017). Comparable approaches have produced agent models that achieve good outcomes in game-theoretical tasks, including general-sum games and complex social dilemmas, by emphasizing cooperation (Lerer and Peysakhovich 2017) and prosociality that takes into account the other's rewards (Peysakhovich and Lerer 2017b), or by conditioning behavior based solely on reciprocity (i.e., cooperating only with opponents that reciprocate cooperation) (Peysakhovich and Lerer 2017a).

Other recent studies have begun to address the emergence of social norms and conventions in multi-agent scenarios (Freire et al. 2020a; Köster et al. 2020). Notably, agents endowed only with model-free RL algorithms were able collectively to follow conventions. Such model-free RL agents do not rely on a model of the task to support reasoning and planning, they learn habits by trial and error. The fact that they can converge toward a common convention might imply that some conventions can be supported by habitual cognition instead of deliberate model-based planning.

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

The above examples ignore one fundamental constraint faced by embodied agents: sensory states must be derived from one's own sensors from an egocentric perspective (Prescott et al. 2018). The majority of agents in the studies mentioned above gather their sensory data from a third-person perspective. They are trained using raw pixel data taken from a computer screen, which can be either completely (Lerer and Peysakhovich 2017; Peysakhovich and Lerer 2017b; Tampuu et al. 2017) or partially observable (Leibo et al. 2017; Perolat et al. 2017). Another limitation present in most of the current research in cooperative AI and multi-agent RL is that they rely on grid-like or discrete environments (Lanctot et al. 2017; Leibo et al. 2017; Perolat et al. 2017; Peysakhovich and Lerer 2017a, b). Although this is an improvement over many classical matrix-form games, insofar as it provides a spatial and temporal dimension (i.e., approximate situatedness), it still lacks the continuous time properties of real-world interactions. Even in the few cases where the coordination task is modeled in real time (Tampuu et al. 2017) and the agents are situated, the aforementioned approaches do not consider lower-level sensorimotor control loops bootstrapping learning at higher levels of a cognitive architecture and the integration of model-free and model-based in unified cognitive architectures. Hence, the generalization of collaborative AI models to real-world tasks and agents remains to be established.

In addition, most of the work developed in this field pursues algorithmic specialists or models that are tested in one single task or environment (Freire et al. 2018; Lillicrap et al. 2019; Perolat et al. 2017; Peysakhovich and Lerer 2017b). This raises a fundamental question about how these models generalize to a more generic or diverse set of problems. At this point, this approach does not readily enable us to extract principles and mechanisms or to unravel the dynamics that underlie human collaboration and convention formation (Freire et al. 2020a; Verschure et al. 2014).

Agents Modeling Other Agents

The discussion above on biological cooperation highlighted the notion of virtualization and mentalizing. A collaborating agent requires a model of other agents, or ToM, with which it can potentially collaborate to align goals, task models, intentions, and actions. Within cooperative AI studies, notable approaches modeling various aspects of ToM already exist, particularly, those based on artificial neural networks (Schmidhuber 2015) and machine ToM (Rabinowitz et al. 2018). These approaches, however, are built on black-box optimization (BBO) algorithms, which hinders our understanding of these models at the mechanistic level. Although BBO algorithms can approximate any complex function, one cannot use them to decipher specific mechanisms and information structures that may underlie ToM in these systems. Therefore, once an objective function or reward heuristic has been set, it is difficult to differentiate the overfitting of sensorimotor couplings that solve the task from a

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

functional ToM that might apply to any social scenario. This creates a paradox: to interpret the mental states of the other, the observant AI agent must rely on an uninterpretable algorithm. This problem is intrinsic to connectionism and was noted in the analysis of its reincarnation in the 1980s (Massaro 1990). Currently, it is the heart of the so-called interpretability crisis in AI and the drive toward "explainable AI" (for a review see Doshi-Velez and Kim 2017; Guidotti et al. 2018). Indeed, for this reason, cognitive neuroscience has converged toward dual-process theories of ToM, which rely on an active interpreter (Gazzaniga 2016; Kahneman 2012).

Another approach toward ToM builds on hierarchical Bayesian inference, which capitalizes on the clustering of social cues (Baker et al. 2011). These methods are cognitively inspired and assume the existence of a prior "psychology engine" in cognitive agents to process ToM computations (i.e., all knowledge that facilitates ToM is provided a priori while the learning system learns to estimate their statistical relationships). Nevertheless, the remaining challenge for this approach is to explain where these priors come from and how the brute force computations required to run these models might be realized in biological substrates of real-world embodied and situated agents. In other words, this approach is reminiscent of the problems faced by "good old-fashioned" symbolic AI, where operational efficiency collapsed under the weight of the predefined world model and the cost of updating its truth values (McCarthy and Hayes 1981) combined with the biological and psychological implausibility of the prior availability of complete world models, or *the problem of priors* (Verschure and Althaus 2003). Hence, this approach can at best inform us on the manipulation of knowledge in the service of cooperation while its scaling to the real-world is questionable.

Open Challenges and Current Limitations of Cooperative AI

Contemporary research in cooperative AI is beginning to integrate more sophisticated machine learning models with the study of population-level dynamics. It mostly follows, however, a technological AI-oriented agenda that focuses on computer engineering problems rather than understanding human or animal collaboration and its underlying cognitive mechanisms. As such, cooperative AI is not directly concerned with the biological plausibility of its models or with providing explanations for collaboration as it occurs in the real world (Freire et al. 2019). Instead, most contemporary research focuses on challenges such as multi-agent planning, multi-objective optimization, and policy convergence and related issues with strong roots in traditional game theory and models from economics, which assume the optimal *H. economicus*. Moreover, in many cases, the focus is on validating algorithmic specialists—virtual agents that make use of one single learning algorithm to solve one specific benchmark (Fujimoto and Pedersen 2021).

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

Open challenges in the field of cooperative multi-agent systems include modeling fully embodied agents that operate within only partially observable environments and are able to learn flexibly across tasks (including meta-learning) when interacting with multiple types of other agents. For an extensive review on the open challenges of cooperative AI, see Dafoe et al. (2021). We note that these challenges have been relevant for the whole of AI since its inception in the 1950s, which reflects negatively both on overall progress and the time constants of memory in the field. Indeed, Allen Newell, one of the founding fathers of AI, defined general intelligence as the ability to make anything a task, while considering social interaction as one of the key benchmarks for theories of cognition (for a review, see Verschure 2023).

From Algorithmic Specialists to Embodied Cognitive Architectures

To understand and emulate biological generalists, we have to shift focus away from algorithmic specialists toward theories and models that can help us understand how different mechanisms of perception, cognition, learning, and control are integrated into one cognitive architecture. This step is challenging and thus far barely explored, yet essential for progress both in terms of generating artificial collaboration as well as in advancing our understanding of biological forms of it in the context of the extended evolutionary synthesis.

From the perspective of the cognitive architecture that underpins social interaction and collaboration, a ten-year research effort in social robotics, supported by the European Commission, is addressing this challenge in the following coupled projects:

- Experimental Functional Android Assistant (EFAA 2023)
- Expressive Agents for Symbiotic Education and Learning (EASEL 2023)
- What You Say Is What You Did (WYSIWYD 2023)
- Distributed Adaptive Control (CDAC 2023)

These research efforts advance a general-purpose cognitive architecture for humanoid robots (DACh) that can be deployed in a range of dyadic collaborative tasks, such as the acquisition of language or teaching. It is also able to engage in a range of collaborative scenarios (Fischer et al. 2018; Lallee et al. 2015). Below, we provide a short summary of the development of DACh to elucidate some key lessons with respect to the understanding, shaping, and creation of collaboration.

The Distributed Adaptive Control Perspective

Distributed Adaptive Control (DAC) is a theory of the design principles that underlie the mind, brain, body nexus (MBBN). It seeks to explain goal-oriented action in physical and social environments and shows how action can

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

result from a control architecture that is organized in four layers—the somatic (the body), reactive, adaptive, and contextual (Figure 5.1)—with tight coupling within and between each layer (Verschure 2012; Verschure et al. 2014). Across these control layers, a columnar organization exists which, at every level of the hierarchy, processes states of the world, the self, and action, where the latter mediates between the former two through the environment. As we move up the DAC hierarchy, states on which operations are performed become more virtualized from the analog signals picked up by the senses serving flexibility in perception, memory, and cognition. In contrast, lower layers are based on priors that allow for rapid yet rigid responses. Thus, the architecture dynamically balances these layers to resolve distinct trade-offs, such as speed and robustness (i.e., reactive control), versus flexibility in problem solving (contextual control). DAC is a theory of brain organization and has been applied to the realization of synthetic collaboration between machines (i.e., humanoid robots) and humans.

The reactive layer can be seen as a model of the evolutionary ancient core behavior systems (CBS) of the brainstem (Merker 2005), which drives behavior and learning based on genetic priors. The adaptive layer facilitates the learning of the perception and action state space, allowing adaptation to an unpredictable world. The contextual layer builds on these acquired representations (virtualizations), affording the construction of goal-oriented policies and their subsequent compression into habits. The contextual layer serves further virtualization through self-monitoring and autobiographical memory (i.e., metacognition, building abstract models of the self, "the other," and the tasks in which it finds itself). To realize this cognitive bootstrapping, DAC includes several learning and memory systems. The adaptive layer is defined as a model of Pavlovian learning, which is constrained by reinforcers detected by the reactive layer. In this way, the adaptive layer acquires states of the world, semantic memory, their association with the agent's actions, and is further expanded through integrative episodic memory. The contextual layer operates on these memory systems (e.g., allocentric maps, and individual items and events stored in semantic memory) to build action policies by constructing sequential representations predicated on goals and values. These policies are retained in long-term memory and operated on in working memory. At every operating cycle, the contextual layer deliberates on which action to emit, by optimizing the expected utility of active policies weighted against the relevance of new perceptual and memory states. Effective policies from this model-based process are compressed into model-free habits and stored in procedural memory to settle the speed-flexibility trade-off.

Biologically grounded cognitive architectures, such as DAC, specify the basic processes that underlie intelligent behavior by including the specific processes of arbitration, representational formats, and conflict resolution that go beyond the implementation of a learning algorithm. A distinguishing feature of the DAC architecture is that each layer of the control system is an integral part

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

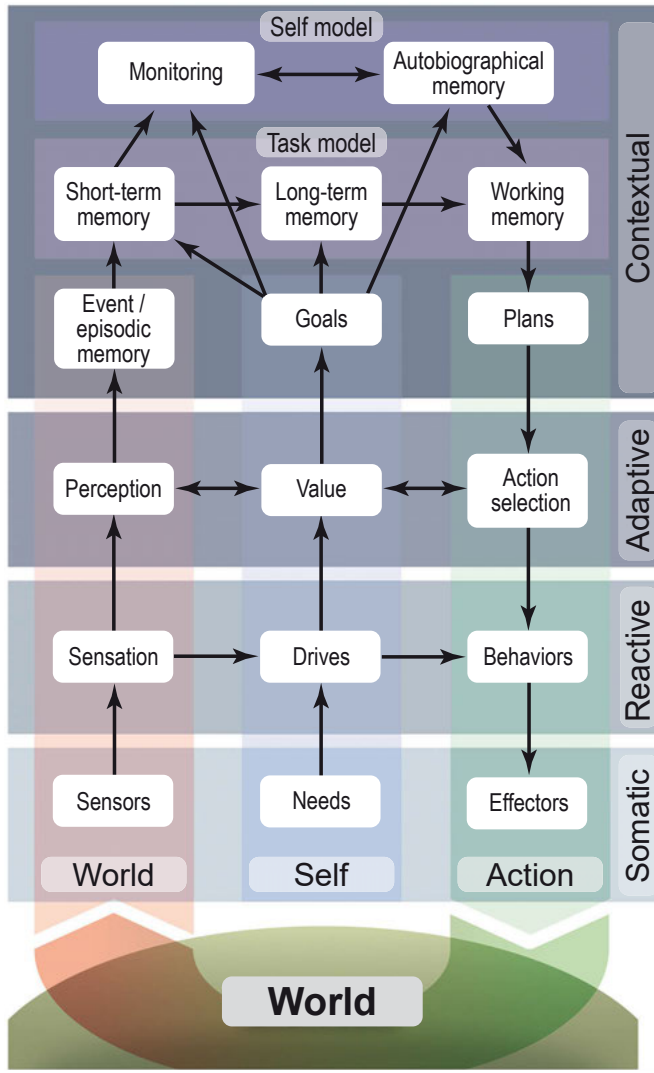

Figure 5.1 A highly abstract representation of the Distributed Adaptive Control (DAC) cognitive architecture showing its main processes (boxes) and dominant information flows (arrows). DAC is organized in four layers (bottom to top): somatic, reactive, adaptive, contextual. Across these layers, three functional columns of organization are distinguished (left to right): exteroception, the sensation, representation, and modeling of the external world (red); interoception, detecting and signaling states derived from the embodied self, from needs and drives to values and goals (blue); and action, which establishes the interface between self and the world (green). The arrows show the primary flow of information, mapping exo- and endo-sensing into action, defining a continuous interaction loop with the external world. Image adapted from Verschure et al. (2014) (Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/).

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

of a larger system. In contrast with standard RL approaches, which focus on the application of just one learning algorithm, the DAC architecture emphasizes integration and complementarity across components. DAC's perspective emphasizes that humans are more than simple stimulus-response automata driven by Thorndike's Law of Effect (Thorndike 1927), instantiated by the reactive and adaptive layers. Rather, behavior is goal-directed and the outcome of the adaptive and volitional human mind/brain interacting with its environment. We argue that the application of biologically informed cognitive architectures such as DAC to the study of collaboration both addresses the limitations of the current multi-agent RL perspective and articulates the key challenges in understanding and synthesizing collaboration.

In the development of an embodied social cognitive architecture, the minimal unit is the dyad of two agents. Understanding this primitive unit will help us to identify the necessary conditions for collaboration. Collaboration depends on the alignment of the goals of the agents involved. Hence, one of the first challenges to generate effective and aligned actions in a dyad is to estimate the mental state of "the other." This raises the question of what the reference is for this interpretation. The standard contemporary AI approach would be to train a network on many prelabeled examples of surface features, such as posture, facial expression, or prosody and their associated mental states. This approach has only shown limited success converging on simple shortcuts rather than ToM (Aru et al. 2023).

An alternative route toward ToM is provided by the phenomenology of Edmund Husserl, Martin Heidegger, and Maurice Merleau-Ponty and its modern incarnation in the embodied cognition movement. DAC has built on the so-called apperception or "other like self" model of Merleau-Ponty, taking into account that "self" is a multiscale notion from the physically instantiated ecological self to the constructed narrative self (Neisser and Fivush 1994) matching the multiple layers of the DAC architecture. Hence, this embodied cognition perspective on ToM proposes that it is grounded in the sub-architecture of the perceptual, emotional, cognitive, and motor substrates that define and represent the self. Direct evidence of this hypothesis was found through the discovery of the so-called mirror neurons by Giacomo Rizzolatti and his colleagues in the early 1990s. Several decades of subsequent research have produced strong evidence that information of the other is broadly mapped onto brain networks that encode states of the self, including areas involved in perception, emotion, cognition, and motor control in a range of mammalian species (for a review, see Bonini et al. 2022).

With "self" as the frame of reference for the representation of the other, the question is: What ontological commitments does the observer need to make to track the states of other agents? The neurophysiology of "other like self" shows that ToM results from multimodal internal models rather than discrete and fragmented representations. DAC proposes that the model underlying ToM can be defined in a six-dimensional conceptual space, known as H5W,

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

comprised of action (how), motivation (why), objects (what), space (where), time (when), and agents (who) (Lallee et al. 2015). We can consider the DAC H5W framework as a minimal situation model in which collaboration can be described, communicated, and modulated. It needs to be emphasized that H5W proposes three priors within which experience is constructed: space, time, and intention (Verschure 2016b). Having established an ontological frame in which social interaction and collaboration can be modeled (or virtualized) by an agent following the "other like self" and H5W model, the question is: How is collaboration initiated, bootstrapped, and maintained? This requires at least three additional ingredients:

1. The machine must be recognizable to its peers as a potential social entity and thus collaborator.
2. It must proactively engage with its peers to create and contribute to collaboration.
3. It must be able to optimize alignment of goals and actions through communication.

In a series of experiments, these aspects were addressed using dyadic human–robot interaction (HRI) paradigms, building on the humanoid robot iCub (Metta et al. 2010). In a first set of experiments, gaze, eye contact, and the expression of basic emotions (e.g., happiness, sadness, disgust, anger, fear, and surprise) were demonstrated to play an essential role in the psychological validity or social salience of HRI (Lallee et al. 2015).

Given that an agent can be recognized as a potential collaborator, the question becomes: How can the actual collaboration be initiated, maintained, and bootstrapped? DACh assumes that at the level of the most primitive reactive layer (i.e., the CBS), one of the intrinsic motivations is to act and interact with social peers. This is based on the notion of "play" as a fundamental behavioral drive, advanced by Panksepp et al. (1984), and the notion of the "human interaction engine," advanced by Levinson (2006). This social CBS includes capabilities for shared attention, pointing, curiosity, reciprocity, turn-taking, and knowledge sharing (Moulin-Frier et al. 2018). As a result of the cognitive bootstrapping in DACh, the agent can acquire goal-oriented behavioral policies that now also pertain to social interactions and collaborative tasks. To ensure alignment in the collaboration, the DACh agent must verbally report on its behavioral policies and experiences stored in its autobiographical memory, which defines DACh's narrative self. The memory systems of DACh allow it to project future actions and reactions, when future states of the world are expected to resemble those encountered in the past. In various HRI scenarios, DACh was validated: the robot had to learn concepts related to its own body, its environment, and other agents in a proactive manner and to express those concepts in the structuring of collaborative goal-oriented behavior. To achieve this, DACh had (a) to balance its epistemic needs of learning about the task and

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

the other, using and expressing its knowledge, and (b) to maintain the coherence and continuity of the interaction.

An intrinsic problem of DACh is credit assignment in multi-agent scenarios. If all external agents are mapped onto internal predictive models that mediate between perception and (social) action, how can these models be further calibrated given the feedback DACh receives? For instance, agents might not behave as expected, inconsistencies in alignment may occur, and specific social norms and conventions need to be considered. The DAC theory proposes that this calibration is the specific function of consciousness; that is, the unification of parallel virtualization processes into one super model that allows compression of parallel models into a unified scene, its valuation, and the adjustment of the underlying models. In this respect, DAC hypothesizes that functional primitives underlying collaboration, including proto-consciousness, emerged during the Cambrian explosion 560 million years ago (Verschure 2016b). Although we cannot second guess evolution, collaboration is a highly conserved feature of biological systems and not a recent invention. Indeed, the earliest genetic changes that gave rise to eusocial animals are at least 110 million years old (Nowak et al. 2010).

Bringing It All Together

A Challenge That Requires Unification

As seen in the brief review of synthetic collaboration in the previous section, current synthetic approaches to studying collaboration are less integrated between themselves than the empirical work on the natural sciences. Although significant progress has been made on each front in parallel, it has happened mostly in isolation, without significant interaction between fields. Building truly autonomous synthetic communities of embodied artificial intelligent agents will require the integration of insights coming from the extended evolutionary synthesis, agent-based modeling, computational neuroscience, cooperative AI, and social robotics. We believe that one of the most promising research avenues in the multilevel study of collaboration will come precisely from the combination of *complex cognitive architectures* at the individual level (such as DACh), with *complex multi-agent environments* at the population level (like the ones currently studied in collaborative AI).

In addition, the extended evolutionary synthesis and its perspective on complex phenomena, such as collaboration, need to incorporate synthetic methods. These methods, as exemplified in the DACh architecture, enable artificial systems to render intrinsically inaccessible first-person states (e.g., the content of memory, ToM, and goals) into third-person accessible constructs.

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

Back to Dusk

Although it still lives in the realm of science fiction, a challenge of the scale and ambition of the construction of the Dusk colony could drive the field forward because it necessitates an intense and concrete interaction between empirical, theoretical, engineering, and computational approaches. The construction of such a colony will require the development of a truly autonomous robot society that is able to self-organize to achieve such a goal without direct human supervision. Achieving this goal implies a lot of trial and error as well as a great deal of imagination. To be prepared for long-term survival, such a synthetic community should be able to recycle the materials of its malfunctioning members and build completely new ones when required. Such "newborn" machines could go through a process of morphological and cognitive development similar to biological systems. Such a developmental process cannot be understood without reference to the social environment in which the agent is embedded. This developmental approach to AI, already advanced by Alan Turing (1950), should be combined with the learning of cultural practices of the community through a process of social interaction and scaffolding as proposed by Lev Vygotsky (1978). This Turing–Vygotskian socio-developmental model of the development of communities of collaborative synthetic agents will also be useful to understand how the process of development and socialization interacts with the cognitive, behavioral, and ecological dimensions implicated in the study of collaboration. The DACh model described above could be considered a first step in this direction.

An alternative approach would be, for instance, to rely on a robot extended mind or a "robot cloud." Back to Dusk, we could imagine how every new robotic exemplar would automatically upload all its perceptual and cognitive protocols from a collective cognition database containing all knowledge the robot colony has generated through its existence. Several simulations of this strategy were tested before launching the final mission to Mars. However, one reason why these Dusk simulations failed was that the robot cloud assumed that the future is fully predictable, and no cognitive scaffolding was required. Its crucial mistake: forgetting that other agents are the main source of unpredictability in the world. Another simulation of Dusk failed due to a sort of *value function hacking* by a rogue group of robots. In this iteration, a group of robots that were in the workshops building and servicing worker robots, adjusted the value functions of the latter, resulting in more ore being mined than necessary. The ore surplus was then used by the "hacker" robots to obtain services from other robot castes so as to reduce their own existential risk. Once this first example of robot corruption was discovered, value function hacking was prevented by adding additional security protocols through rapid optimization. Unfortunately, this redesign was not sufficiently tested; mining operations, previously a task of humans, ground to a halt; the hacker robots turned to outright sabotage; and the colony collapsed before humans could

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

interfere. This failed simulation of Dusk became known as *capitalization on the commons*; that is, the real or imagined objects in the world that pertained to the value function of a group of collaborating agents was identified and thus could be instrumentalized through a parasitic attack that drove up the cost of access to the commons.

The Shape of Things to Come

The rise and fall of the Dusk colony happens in a hypothetical future, yet science fiction stories, regardless of their futuristic envelopes, usually embody a warning about the present. That lesson not only involves reflecting on what we can do today to avoid a disturbing future; it also involves realizing which elements of such a future are already here with us today, and critically reflecting on what we can do about it.

From Cybernetics to Cyberethics

The possibility of an open-ended process taking place in the development of intelligent artificial agents and their role within human societies can lead directly to concerns about predictability and control. Without predictability and control in the design of collaborative processes between humans and machines, we will enter the realm of the "unknown unknowns" of the morphospace of collaboration, which clearly have worrying ethical implications due to unforeseen (and potentially catastrophic) consequences. As the fictional complex systems theorist Ian Malcom pointed out in the film *Jurassic Park*, "scientists were so preoccupied with whether or not they *could*, that they did not stop to think whether they *should*." Indeed, before we engineer and develop increasingly competent and intelligent artificial agents, like the protagonists of the Dusk colony, we should first address the study of the ethics of machines.

Efforts in this direction are currently taking place within the field of *AI ethics*—a multidisciplinary field that tries to address the current ethical issues regarding the growing impact that artificial intelligence and robotics have in our society (Coeckelbergh 2020). The field addresses how to implement moral decision making in artificial agents as well as the ethical consequences of introducing artificial agents in human environments and how to manage issues that arise. Examples of this line of research cover issues such as the study of algorithmic biases (Kordzadeh and Ghasemaghaei 2022), algorithmic transparency (Brauneis and Goodman 2018), and algorithmic accountability (Kemper and Kolkman 2019; Wieringa 2020).

An even more novel research framework has recently been proposed under the name of *computational ethics* (Awad et al. 2022), which aims to bridge a gap on the AI ethics research program by incorporating insights from cognitive science. In short, computational ethics specifies how the ethical challenges of

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

AI can be partially addressed by incorporating the study of human moral decision making. The perspective followed by the computational ethics framework is very much aligned with the synthetic spirit of this chapter: it attempts to unify both empirical and computational approaches under one single research program. More concretely, it proposes to formalize our current understanding of human ethics computationally so it can be applied to the development of machine ethics. Interestingly, the study of the underlying computational principles governing behavior in humans and machines has its roots in the field of cybernetics. In honor of such a research tradition, we think that a more fitting name to the study of computational ethics should be *cybernethics* or *cyberethics*.

Collaborative Cybernetics: Toward a Multilevel Cybernetic Approach to the Study of Collaboration

In the eponymous book that gave birth to the field, Norbert Wiener (1948) defines cybernetics as the study of control and communication in the animal and the machine. The word cybernetics comes from the Greek words κυβερνάω (kybernáō), which means "to steer, navigate, or govern," κυβερνητική (kybernētiké), which means "governance," and κυβερνήτης (kybernḗtēs), which refers to the governor or helmsperson of a ship. Since the notion of feedback or circular causality is central to the study of cybernetics, the name of this field was coined based on an example of steering a ship. To reach its destination across the seas, a helmsperson needs to maintain a steady course by continuously adjusting the steering of the ship in response to the waves and changing winds. In a way, a collaborative task is like a ship in which each member must work together to steer it in the right direction. Thus, every collaborative effort requires more than one helmsperson.

In the spirit of this Forum, we advocate for synthesis across fields to help us better understand what collaboration means. The final goal is to develop a cross-disciplinary, multilevel theoretical view of collaboration. This emergent field should aim to bring together the study of collaboration in both biological and artificial agents, with the goal of extracting general rules to understand and develop collaborative multi-agent systems, both hybrid and unblended (i.e., human–human, human–machine, and machine–machine collaboration). We propose to name it "collaborative cybernetics"—the study of coupled, interdependent, goal-oriented systems that share a common goal.

We believe that the definition and development of such a research program is more relevant than ever. As humanity has become more interconnected and globalized, human societies are more interdependent on each other. The consequences of such globalization feed back onto the environment, affecting not only our species but the totality of Earth's ecosystems. In his famous speech to the United Nations in 1965, Adlai Stevenson said:

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

> We travel together, passengers on a little spaceship, dependent on its vulnerable reserves of air and soil; all committed for our safety to its security and peace; preserved from annihilation only by the care, the work, and, I will say, the love we give our fragile craft. We cannot maintain it half fortunate, half miserable, half confident, half despairing, half slave—to the ancient enemies of man—half free in a liberation of resources undreamed of until this day. No craft, no crew can travel safely with such vast contradictions. On their resolution depends the survival of us all.

Viewing our planet as a shared spaceship helps us understand how deeply linked our destinies are to that of our ship. Its correct maintenance and survival are responsibilities that each one of us must bear, as we are all helmspersons of Spaceship Earth. Unfortunately, an instruction manual did not come with it, as Buckminster Fuller once famously remarked. It is time to start making one, for if we want to steer the ship toward a safe destination, we must learn to steer it *together*.

From "The Nature and Dynamics of Collaboration,"
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144

From “The Nature and Dynamics of Collaboration,”
edited by Paul F. M. J. Verschure et al. Strüngmann Forum Reports, vol. 33,
Julia R. Lupp, series editor. Cambridge, MA: MIT Press. ISBN 9780262548144