Exploring Ethical and Conceptual Foundations of Human-Centred Symbiosis with Artificial Intelligence

Antonio Carnevale [1,2], Antonio Lombardi [3] and Francesca A. Lisi [4]

[1] *University of Bari Aldo Moro, DIRIUM Department, Palazzo Ateneo, Piazza Umberto I, 70121, Bari, Italy*
[2] *DEXAI Srls - Artificial Ethics, via Macedonia 73, Rome, Italy*
[3] *University of Bari Aldo Moro, DIRIUM Department, Palazzo Ateneo, Piazza Umberto I, 70121, Bari, Italy*
[4] *University of Bari Aldo Moro, Dipartimento di Informatica, Via E. Orabona 4, 70125 Bari, Italy*

Abstract

Symbiosis between humans and AI is a two-way relationship that poses several unprecedented challenges not only from the technological viewpoint but also as regards foundational AI research. In this paper we address some philosophical questions about the nature of this symbiosis and argue that a human-centred approach is needed to design symbiotic AI systems in order to ensure their ethical acceptability.

Keywords

Symbiotic Artificial Intelligence (SAI), Human-Centred Computing, AI Ethics, Machine Ethics, Philosophy of Technology

1. Introduction: Overview on Symbiotic Artificial Intelligence (SAI)

As Artificial Intelligence (AI) systems are becoming a part of our daily lives, how to improve the current shortcomings and limitations in human-machine collaboration becomes a question more urgent than ever. The primary challenge of an AI agent functioning alone is how effectively and flawlessly it achieves its goal. However, in a team of humans and AI assisting each other for a common goal, the challenges are not limited to the goal itself, since the AI system should have the ability to reason about the human's actions while considering their mental models. *Symbiotic AI* (also known as human-AI symbiosis) promises to boost human-machine collaboration and sociotechnical teaming, with mutually beneficial relationships, by augmenting (and valuing) human cognitive abilities rather than replacing them [1]. Sociotechnical teaming refers to the collaborative partnership between humans and machines within a broader social and technological context, where the focus is not on a substantial peer-to-peer relationship but on integrating technology into human-centric processes and systems. In this context, symbiosis involves humans and machines working together as a cohesive unit, each playing a specific role and contributing to the team's overall performance. Humans provide the cognitive and emotional capabilities necessary for creativity, empathy, ethical decision-making, and adaptability. On the other hand, machines offer computational power, data processing, and automation capabilities that can handle repetitive and data-intensive tasks efficiently. This sociotechnical teaming approach is prevalent in various domains, including manufacturing, healthcare, finance, and education. For example, doctors and diagnostic AI systems collaborate in healthcare to provide more accurate and timely patient care. Human workers collaborate with robots in manufacturing to improve production efficiency and product quality. This collaborative dynamic optimises the strengths of both parties, enhancing

[1]AIxIA Workshop on Bias, Ethical AI, Explainability and the Role of Logic and Logic Programming, November 06–09, 2023, Rome, Italy
EMAIL: antonio.carnevale@uniba.it (A. 1); antonio.lombardi@uniba.it (A. 2); francesca.lisi@uniba.it (A. 3)
ORCID: 0000-0003-2538-5579 (A. 1); 0000-0003-1803-5423 (A. 2); 0000-0001-5414-5844 (A. 3)

productivity and problem-solving while focusing on human values, ethics, and decision-making. It's not a strict peer-to-peer relationship but a cohesive teaming strategy that leverages the strengths of both humans and machines to achieve superior results [2, 3].

Fostering human-AI symbiosis requires designing AI systems according to a *human-centred approach*, as developed within the Human-Computer Interaction (HCI) community. These systems take advantage of computer features, such as powerful algorithms, big data management, and advanced sensors. However, they also consider the needs, desires, emotions, intentions, actions, etc., of people who are going to use them, providing high levels of automation along with human control. At the same time, it is also fundamental that humans understand and trust robust and stable AI systems that use important knowledge components and make decisions explainable, also in conditions of noise, uncertainty, and small perturbations. The focus is therefore on the design of new interaction paradigms that can amplify, augment, and enhance human performance, in ways that make systems reliable, safe, and trustworthy. This is the main question of the foundational research done by the University of Bari (together with INFN) within the NRPP-funded project Future AI Research (FAIR).[2] It is articulated into several subquestions, among which we here address the following: How to improve the understandability, acceptability, and sustainability of SAI systems? In particular, *acceptability* is the subject of research for our investigation within a dedicated work package in the project we are involved in. It involves *value alignment* between AI and humans. It is related to understanding AI decisions, the algorithmic bias, the respect of privacy policies for data collected by AI systems, the struggle between security ensured by AI systems and fundamental freedoms, the mitigation of possible safety and health risks. In FAIR, studies on the acceptability of SAI adopt an interdisciplinary approach involving researchers in AI, Law, and Philosophy. The goal is twofold: (i) to investigate the respect and exercise of fundamental rights of humans supported by SAI (e.g., legal rules in algorithmic decisions, AI and responsibility, robotic subjectivity, workers' rights, and taxpayer's rights); (ii) to study the epistemological and ethical aspects of SAI. In this paper, we contribute to the achievement of (ii) by exploring the foundations of human-centred symbiosis with AI from the viewpoint of philosophers.

The structure of the paper is the following. In Section 2 we address some foundational and philosophical aspects of SAI, by analysing first the relationship between life and artefact (Section 2.1) and then the one between intelligence and symbiosis (Section 2.2). In Section 3 we investigate the factors that might enable a symbiosis with AI systems, by exploiting synergies between the fields of Ethics and Human-centred Computing. First, we analyse ethical dilemmas arising from the application of human-centred approaches to SAI (Section 3.1). Then, we introduce our conceptual model for mapping and assessing symbiosis in socio-technical systems (Section 3.2). Additionally, we offer thematic areas for potential policy recommendations to facilitate workshop discussions. Notably, these recommendations arise from our method's conceptualization phase, not from the assessment results (Section 3.3). In Section 4 we conclude the paper with final remarks and directions for future work.

2. Foundational and Philosophical Aspects of SAI

The term 'Symbiotic Artificial Intelligence' adds to an already problematic binomial – that of 'Artificial Intelligence' – an adjective that further complicates the picture, but which seems to constitute one of the very latest frontiers of AI research. In 'symbiosis', it is implicit that these are two or more lives that coexist in order to cooperate (or at least to benefit from coexistence with the other). Therefore, a truly Symbiotic AI should envisage an interaction between man and machine that is no longer that between a controller and a controlled, but rather a 'biunivocal' one between two potentially 'equal' agents in the decision-making process, and in which one can influence the other. Two intelligent agents, then, if perhaps not in the same way. The 'ex-controller' – the man – engages in a relationship in which it is the machine itself that can take over the control functions by modifying, correcting, intervening in its choices in real time and in the direction of a common 'goal', based on a broader and more precise knowledge of the data and procedures required to perform the task. In this regard, it should be noted that no matter how asymmetrical this kind of relationship

[2] https://future-ai-research.it/

between (intelligent) lives may be, there is always a residue of control that one party exerts over the other, and vice versa. Otherwise the symbiotic relationship would cease to exist altogether. However much, for example, between human beings and *Escherichia coli* there is a symbiotic relationship strongly unbalanced on the former, the bacterium maintains a certain amount of biological autonomy that, on the one hand, promotes the well-being of the host through a kind of non-mental wisdom while, on the other hand, it always runs the risk of giving rise to a pathological relationship. We do not have complete control over *Escherichia coli*, although we rely on its expertise for the success of our metabolic activity.

This type of HCI – which we define as symbiotic – is also different from that of so-called 'third-order' technologies [4], *i.e.* where one technology controls another technology by means of an intermediate technology and the processes are still automatic: although an important dose of artificial intelligence 'autonomy' is observed here, man always plays the role of ultimate supervisor and the machine's capacity for initiative is almost nil.

SAI, on the other hand, could in principle also operate differently from its symbiont, perhaps because the latter has made incorrect assessments. At the same time, this is possible because, conversely, man relies on artificial intelligence, anticipating from the outset that his cooperation with it will take place according to this mode of interaction.

How is it possible to think of a SAI? What kind of technology will it be? Are there already forms of human-machine interaction that foreshadow their relationship in a symbiotic sense? For instance, Kai-Fu Lee and Chen Qiufan [5] imagine a not-too-distant future in which smartphone apps will have so much real-time data about us and our daily actions that they will prompt us to change our behaviour or make certain decisions in contexts assessed by a complex system of apps interacting with each other and with us (through our devices). Is it, then, just a question of the availability of personal and environmental data or the potential of the technologies at our disposal that can enable the symbiosis between natural intelligence and artificial intelligence? Or is there something more, or different?

Certainly, to think of an AI means to imagine the possibility for the machine or software to be able not only to learn and respond 'live' (this to some extent already happens with the IoT, for example), but also to be able to learn regularities from our behaviours, intervening in real time, and with a certain amount of 'authority': in the sense that it is potentially capable of modifying human behaviour that, in absence of its artificial counterpart, would be exercised otherwise. Symbiosis in this sense provides for a kind of peer relationship, in which two lives coexist and learn from each other: the machine learns by observing me, and I learn new information that the machine provides from what it has learned by observing me.

"While the computer's role has shifted from being a passive 'executor' to an active learner, the role of the human is still that of an actively involved teacher, because it is the humans that create clean, consumable data for the computer to analyse and hence train itself. However, what if there was a way in which the computer's role remains that of an active learner, but we humans can passively sit back and go on with our lives, while the computer learns from our actions?" [6, p. 4].

Here, it is no longer a question of 'teaching' machines what to do, but of learning from them: not only in the indirect sense that this can always happen (chatting with Chat-GPT on a subject we know little about, for example), but in the sense that they can give us directives on what to do in order to do it better, and achieve the desired end. Much scientific literature calls this type of HCI "symbiotic," but to what extent is this correct? What meaning of symbiosis is involved in this application of the term to AI? And with what consequences?

The purpose of this Section is to try to give some answers to these questions, which can be summarised as follows: how can a Symbiotic Artificial Intelligence be defined and in what ways, if at all, is it possible? These foundational questions will allow us to address whether and to what extent a symbiotic type of AI is virtually compatible with a human-centred approach.

2.1. Life + Artefact: How is Symbiosis Possible for Machines?

The first thing to focus on regarding the possibility of SAI from a foundational point of view is the relationship between *life* and *artefact*. These are two elements that are traditionally opposite. A living being is one thing, an artificial being is another: here, however, they must somehow be together. Traditionally we trace the great distinction between entities of nature (and therefore living entities) and artificial entities back to the second book of Aristotle's *Physics*. But this goes all the way to Darwin, who still distinguishes between two great kingdoms, or two great systems of laws in the world, which are the organic and the inorganic [7]. So, at first glance it would seem to us that these are two distinct systems of reality, but in our acronym they come together.

According to this traditional paradigm of the life sciences that goes from Aristotle up to modern biology, an artefact, or a machine (an entity that is a product of *techne*, according to Aristotle's own definition) is not something alive. What does it mean to be alive for Aristotle? In *De anima* Aristotle says that to live means to have the capacity to nourish, grow and decay [8, p. 22]. These functions are all united under the sign of movement, specifically spontaneous movement. For Aristotle and then throughout the long course of Western biology, what distinguishes a living thing from an artefact is the ability to move spontaneously. The machine is not endowed with life, it is not "animate" (ensouled) because it does not possess an autonomous principle of movement. It is we who move the machine, whereas a living thing, an insect, or an animal, such as plants themselves (which have a nonlocal motion, but a motion of growth), move autonomously. We do not drive them.

The modern era, especially with the rise of the mechanistic model, will reject such a view of life by coming to conceive even living beings as machines whose movement is always the result of a complex chain of external impacts (think of Descartes, Hobbes, and Spinoza). But a revival of Aristotelian conceptuality will occur in the nineteenth century, with *Naturphilosophie* and the emergence of an "organicistic" paradigm, which gave birth to modern biology.

This conception, whereby a living being is one that possesses an autonomous principle of movement, a self-finality (*entelechy*), has been somewhat challenged by the emergence of robotics and computer science in the twentieth century. These sciences study the ways in which it is possible to build machines that can attain a certain level of autonomy. Obviously not the autonomy that pertains to living beings in general, and to the human being less than ever. But all artificial intelligence systems, and today especially Large Language Models (LLMs), have shown the capacity for a kind of self-development. Robotics and computer science in some ways constitute an exception to the old distinction between *physis* and *techne*, between life and artefact. Robots and instances of artificial intelligence are a special kind of machines: they confront us with "mechanical" processes that nevertheless exhibit characteristics that somehow have to do with life, and thus potentially also with symbiosis. Maturana and Varela proposed in the twentieth century the famous distinction between *autopoietic* and *allopoietic* machines: the former machines would be the living organisms, that is, machines whose process results in the preservation, reproduction, and development of the machines themselves, while allopoietic machines are machines that are operated and have something other than the machine itself as a result of their process [9].

One might ask: If an artefact (a device or technology that is not made of organic parts) possessed all the characteristics that an organism that is made of organic material also possesses, what would remain different between the two? Would they both be considered living? [10 p.55, 11] Can an artefact become an autopoietic machine? Indeed, this was a route attempted as early as the middle of the last century (e.g., by Von Neumann, with his self-replicating machines) and then in the 1980s in the field of A-Life.

A-Life is based on the assumption that life is something that can also be synthesised digitally: that is, it does not matter what the support is, whether it is made of atoms or bits. The important thing is that this support manifests certain characteristics, precisely including that of self-preservation, self-reproducing, having autonomous movement aimed at its own sake, etc. Examples are numerous (Conway's Game of Life, Polyworld, RepRap, Slugbot etc).

These developments seem to challenge the classical distinction between life and artefact, and this is of utmost importance for the purpose of a foundation of a SAI. Indeed, the latter can be thought of as an AI that virtually has the ability to enter into a "life-to-life" relationship.

Are robots "living" entities? Are artificial intelligences such as ChatGPT entities that, although not "material," exhibit some of the characteristics of life?

In the case of robots Lévy [12] answers that indeed, according to some widely accepted parameters in biology (such as [13]), it is not entirely improbable to consider robots and therefore also AI programs as life forms. Differently thinks de Collibus [14] about LLMs. A number of characteristics that we are spontaneously inclined to assign to living things are also found in ChatGPT: for example, the ability to self-develop and learn, which is a capacity we possess as well as a dog or a fish. Like the latter ChatGPT is able to treasure the data it collects and thus develop greater complexity and better and better solutions in successive attempts to deal with problems. However, these programs lack a fundamental characteristic of the living, which is what Spinoza called *conatus* and Schopenhauer the will to live, that is, the instinct for self-preservation. None of these technologies actually act because they are driven by the will to remain in existence. Perhaps the self-finality that Aristotle first identified as the main characteristic of the living is not fully found in an AI: it is indeed capable of self-growth or self-development, but why does it do so? It does not seem to do so because it aims at its own self-assertion or its own "good." This could be an insuperable hiatus between living beings and machines, and a serious objection to the possibility of applying biological categories such as symbiosis to AI technologies.

2.2. Intelligence and Symbiosis: Three Patterns Based on (Dis)continuity

How do things stand with intelligent life? For the foundation of an SAI, the question is essential, since it is assumed that the two lives that should enter symbiosis are both intelligent lives. And indeed, in the case of the artificial symbiont it might be the case that its life completely coincides with its intelligence. SAI thus allows us to ask important questions about the kind of relationship that exists between life and intelligence.

The problem is extremely difficult, since it depends on what kind of definition of intelligence we start with.

A minimal definition of intelligence as the capacity for organisation has been advocated by various philosophers and scientists, starting from Schelling and Darwin. In this sense, even a mushroom or a worm (or our own organs, as Darwin says) possess a minimal degree of intelligence, since they exhibit a certain amount of organisation aimed at solving problems and adapting to the environment [15 p. 21]. But this is what software, programs, and instances of artificial intelligence also do: they have a basic arrangement that starting from certain inputs found in their own context reconfigure themselves to meet specific demands. From this perspective, which could be called 'homogeneous continuism', it is clear that a SAI is already there and is already with us: our devices, through which we interact with generative AI systems, are already in some sense forms of intelligence with which we live in symbiosis. Clearly, here we are dealing with a definition of intelligence that flattens discontinuities and reduces them to a mere difference in degree: it is enough for an entity to exhibit a certain computational strategy to be considered intelligent (from viruses to humans and AI systems). At this level, a SAI is conceivable first and foremost as a form of integration between human beings and digital tools [16, 17], in a relationship that provides for their increasing autonomy, seamlessness and self-development. A perspective that has begun to be investigated in the *White Papers (2017-2020) on Symbiotic Autonomous Systems* (SAS) [18, 19, 20].

However, there are less inclusive definitions, such as those that restrict the field of intelligence to vertebrates with a minimum of brain activity [21]. In this case, calling that of AI an "intelligent" life would already become less obvious. Floridi [22] for example argues that current models of generative AI do not even reach the degree of intelligence of a sheepdog, and are still as stupid as a dishwasher (see also [23]). If we embrace this idea, it is clear that a SAI could not be configured in the terms of our relationship with certain kinds of technologies (such as a smartphone or ChatGPT) but should rather be imagined as the relationship of a human being with a robot with artificial

intelligence that reaches at least the level of an animal. In this case, a true SAI could be realised only at the exosymbiotic and not endosymbiotic level: that is, wearable or prosthetic AI technologies would probably be excluded from SAI. But the problem is that the realisation of "animal" level AI still remains highly problematic.

In opposition to such weaker forms of "continuism," there are various versions of strong or absolute discontinuism, whereby intelligence is actually a uniquely human capability. To define intelligence as any ability to solve problems or to calculate even the way animals do would be an equivocal way of talking about intelligence, because true intelligence is only our own. Indicative of "true" intelligence would only be capabilities such as universalization, creativity, spontaneity, self-consciousness, emotionality etc.: all characteristics that are the exclusive preserve of human beings. In this case, in order to talk about SAI we would have to approach science fiction or very distant in time perspectives that contemplate the possibility of achieving an AGI, and thus some kind of Singularity. At this level, the foundation of a SAI would be very problematic, and we could use this category only in a metaphorical way: that is, we could talk about some everyday practices of interaction with AI or some forms of HCI only “as if" they were a symbiosis between human being and AI, but in reality, at the foundational level a real SAI would never occur. Therefore, the identification of objective criteria for being able to speak of SAI should be obtained more on the level of a socio-technical constructivism than on the level of an ontological and/or epistemic foundation.

This is the perspective taken in the rest of the paper.

3. Synergizing Ethics and Human-Centred Computing: Paving the Way for a Potential Symbiosis with AI Systems

The present Section embarks on an exploration of the intricate dynamics existing between distinct yet interrelated paradigms. Principally, it elucidates the inherent tension at the nexus of human-machine symbiosis and methodologies aligned with human-centeredness, as progressively mandated or suggested by diverse international entities, prominently exemplified by the EU as documented in the *Ethics guidelines for trustworthy AI* (2019) and the *White Paper on Artificial Intelligence* (2020).

The ensuing discourse unfolds in a tripartite structure, each section contributing distinct layers of comprehension. Section 3.1 dissects the dichotomy between the evolving imperatives of human-centric approaches, underscored by an assortment of international organisations, and the trajectory of SAI systems' advancement. Section 3.2 proposes an innovative conceptual framework – an intellectual scaffold, as it were – for apprehending the symbiosis hypothesis between the human cohort and AI systems. The ultimate tier of our discussion emerges in Section 3.3. Herein, a compendium of strategic policy alternatives takes form, with the central purpose of harmonising the intricate interplay between ethical considerations and the domain of human-centred computing.

3.1. Ethical Dilemmas at the Nexus of Human-Centric Approaches and Symbiotic AI Advancements

As comprehensively expounded in Section 2, the multifaceted concept of symbiosis accommodates diverse definitions and foundational concepts within its ambit when applied to the intricate nexus between humans and machines [1]. These conceptual underpinnings may manifest through an organic lens, suggesting a measure of equivalence between the entities of humans and machines. Alternatively, an interactionist perspective may emerge, emphasising that the crux of symbiosis lies in the realm of behaviours and interactions, independent of the intrinsic nature of the agents involved.

However, regardless of the interpretative avenue pursued, an essential quandary emerges: the notion of symbiosis within the context of artificial intelligence seemingly collides with the ethical pursuit of human-centred AI. The tension resides in the dilemma of reconciling symbiosis – whose

essence is entwined with the realm of machines; entities inherently distinct from humans – with the ethical imperative of constructing systems attuned to human essence (whatever this essence is). In this light, symbiosis assumes the role of an apparent contradiction against the backdrop of human-centeredness.

This contradiction gains further traction and efficacy – transforming into more than mere logical negation – due to the burgeoning emergence of normative guidelines across the global landscape. These guidelines overtly or subtly endorse human-centredness as an ethical cornerstone for the evaluation and standardisation of AI. Notable references in this sphere include *The Montreal Declaration for a Responsible Development of Artificial Intelligence* (2018), *OECD Principles on Artificial Intelligence* (2019), *Singapore's Model AI Governance Framework* (2019), *European Commission's White Paper on Artificial Intelligence* (2020), and U*NESCO's Recommendation on the Ethics of Artificial Intelligence* (2021).

Crossing this intricate confluence, we propose a potential avenue to harmonise the development of SAI systems with the underpinnings of human-centeredness. It involves viewing the manifold semantic dimensions of symbiosis as integral elements of human-centricity. Realising one configuration of SAI over another hinges upon our choices, from machine construction to purposeful production, from autonomous machine learning to delineating the bounds of this machinic self-learning. These choices, particularly those concerning symbiosis, crystallise human-centeredness into a practical embodiment of responsibility. Augmenting the symbiosis-human-centricity nexus is the realisation that symbiosis with technology embodies a discourse of perspective – an intellectual trajectory – rather than an entrenched domain of techno-scientific exploration. This presents SAI systems as akin to what the German sociologist and philosopher Max Weber delineates as an "ideal type". An ideal type functions as an abstract and hypothetical construct used in social science to dissect and comprehend complex social phenomena. It captures essential characteristics while transcending strict realism, serving as a conceptual scaffold for understanding intricate realities. In light of this, the pivotal query shifts from "What does symbiosis signify?" to probe the sociotechnical milieu in which symbiosis is tenable [24, 25, 26, 27, 28]. Rather than scrutinising whether symbiotic machines possess human-like cognition, the focus sways towards comprehending why they appear to exhibit such traits [29]. In lieu of envisioning a future dominated by algorithmic decision-making, our directive should be to design human-centric systems where machine decisions resonate with value-based principles.

3.2. Embarking New Foundational Horizons: Constructing the Assessment of SAI through the Socio-Technical Landscape

To embark upon the conceptual elucidation of a theoretical underpinning for SAI systems from a human-centred standpoint necessitates a foundational premise – an intellectual scaffold, if you will. It is essential to cast aside conventional notions of symbiosis as an intrinsic, preordained state or an aspirational endpoint. Instead, we must reframe symbiosis as an ontological potentiality, a prospect rather than a static reality. This resituating of symbiosis within the domain of possibility underscores its human-centred essence. The symbiosis between human cognition and algorithms exists not as an ineluctable biotechnological or post-humanist outcome but as a plausible discourse – an avenue to articulate the burgeoning socio-technical infrastructure that increasingly interlaces human existence with technological rationality. This socio-technical constructivism offers us different benefits:

- **Flexibility in definition**. Artificial symbiosis is a concept that evolves alongside technological advancements and societal shifts. A socio-technical constructivist approach recognises that the definition of artificial symbiosis can be problematic and should not be fixed in stone. It acknowledges that our understanding of this symbiosis is contingent upon the current state of knowledge, technology, and society. This approach allows for ongoing adaptation and refinement of the definition based on empirical testing and real-world implementation scenarios, ensuring it remains relevant and adaptable to changing circumstances.
- **Anthropocentric value clarification**. The socio-technical constructivist approach acknowledges that the anthropocentric value of artificial intelligence is not absolute but

rather a perspective rooted in human interests and values. It doesn't seek to impose one fixed viewpoint but encourages the exploration of various value systems and their interactions with technology. By recognising the subjectivity of anthropocentrism, this approach facilitates a more inclusive and open discussion about the roles and regulations of human-machine symbiosis. It allows us to incorporate diverse ethical and cultural perspectives within a framework that does not privilege any single viewpoint but recognises the value of coexistence and coevolution between humans and machines.

- **Adaptability to emerging AI-based technologies**. As AI evolves rapidly, a socio-technical constructivist approach can accommodate the emergence of new technologies and their impact on the symbiotic relationship between humans and machines. It enables us to consider and adapt to unforeseen developments, ensuring our understanding and regulations remain relevant and practical. This flexibility is essential in an era where technology is advancing at an unprecedented pace.
- **A cautious approach to unbridled techno-ideology**. Contemporary understanding underscores that technological potency is anything but neutral – it assumes a role in shaping and co-determining societal structures [32, 33, 34, 31]. Thus, embarking on a constructive exploration of the conditions underpinning human-machine symbiosis facilitates a precautionary stance against unchecked techno-ideological fervour, preventing succumbing to facile exuberance.

How can we effectively evaluate a socio-technical system? What criteria should we consider when identifying and assessing the level of symbiosis within the system? Evaluating a socio-technical system is notably more challenging than evaluating an individual technology due to numerous variables, contextual variations, scenarios, and changing user dynamics. In our ongoing research as part of the FAIR project, we are developing a model that thoroughly defines the key components of a given socio-technical system. These components represent the conditions that allow us to detect and subsequently consider the system for further assessments. We have conceptualised these conditions as a continuum, transitioning from a predominantly machine-oriented symbiosis to a more human-centred one. These conditions include Technical Functional Performance, Reality-Characteristic Replication, Human-Machine Interaction Type, and Subjective User Experience. Our approach involves using indicators placed on value scales to (a) measure the socio-technical landscape along each of these axes and (b) provide a general framework for mapping the degree and depth of symbiosis within the socio-technical construct under investigation.. Details are reported in Table 1.

Table 1
Socio-technical conditions table for mapping human-centeredness of SAI

Socio-technical condition	Description	Features
Technical Functional Performance	This condition pertains to the operational efficacy of the SAI system as a technical entity. The system comprises diverse components, each endowed with distinct technical functionalities. Through their orchestrated interplay, the system attains operability and usability.	***Relationship Type***: Machine-to-Machine (M2M) ***Artificial Centrality***: High (++++) ***Human Centrality***: Minimal (+)
Reality-Characteristic Replication	This condition delves into the capacity of the SAI system to emulate specific attributes of (human and non-human) reality through its technical operations. The system's technical prowess facilitates replicating, simulating, and reproducing key real-world features.	***Relationship Type***: Machine-to-Human-to-Machine (M2H2M) ***Artificial Centrality***: Significant (+++) ***Human Centrality:*** Moderate (++)

Human-Machine Interaction Type	This condition addresses the nature of interaction between the human and the machine, irrespective of the subjective experience of the former. It underscores the intricacies of the interaction itself, devoid of the human's subjective engagement.	***Relationship Type***: Human-to-Machine (H2M) ***Artificial Centrality***: Moderate (++) ***Human Centrality:*** Substantial (+++)
Subjective User Experience	This condition centres on the subjective encounter of users within the AI system's ambit. It delves into the depth of experiential engagement individuals undergo while interfacing with the technology.	***Relationship Type***: Human-to-Machine-to-Human (H2M2H) ***Artificial Centrality:*** Minimal (+) ***Human Centrality:*** Profound (+++)

.

3.3. Strategic Policy Avenues to Cultivate Human-Centeredness within SAI Systems

In the nascent stage of our study, we are actively honing the criteria and dimensions essential for the evaluation of symbiotic AI as a socio-technical construct. As we wrap up our contribution, we present a set of policy recommendations. It is important to note that these recommendations do not constitute finalised conclusions based on pending analysis results; instead, they serve as a pertinent policy framework for discussion that we plan to delve into more deeply during our upcoming workshop. We believe that these recommendations lay the groundwork for potential regulatory guidelines that can inform the subsequent assessments in the remainder of our FAIR study. These policy frameworks encompass:: (a) Science and education, (b) Ethics of technology, (c) Design of technology, (d) Sustainable development goals (SDGs), and (e) Citizen participation. The policies are described in Table 2.

Table 2
Human-centeredness policies table

Topic	Strategic policy	Description
(a) Science and education	AI as a distinct science	Promoting scientific uniformity in AI offers benefits: (a) helps data scientists keep up with advancements; (b) shifts focus from business to research; (c) defines testable boundaries for AI's societal impacts.
	Explainable operationalisation of knowledge	Knowledge, including data, grows exponentially. The need is shifting toward mastering skills for knowledge access and its operationalisation.
(b) Ethics of technology	Ethics as a critical inquiry	Even with value-sensitive innovation, technology design assumes development, hindering its assessment for cessation. The freedom to question and decide is crucial; responsible innovation can mask maintaining the status quo [35, 36].
	Proactive ethics	AI ethics faces resistance from some in the field who view it negatively. Embracing ethics involves short-term trade-offs but leads to long-term sustainability. Proactive ethics aids policy with comprehensive approaches, not just bans, offering support and incentives [36].

	From the “what” to the “how”	Scholars urge shifting AI ethics from “what” to “how” [22] - focusing on implementing ethical principles effectively in SAI systems, acknowledging their impact on users, developers, and policymakers, often overlooked aspects of life.
	Ethics and integrity	Translating ethics into best practices faces risks, including ethical shopping, bluewashing, lobbying, dumping, and avoidance, which can hinder even well-intentioned efforts [22]. While not new, these risks take on unique characteristics in AI ethics.
(c) Design of technology	The complexity of human nature	Creating AI for all requires a deep understanding of human behaviour, psychology, and evolving social dynamics. The human-centric approach aligns AI with values, but turning this into specific guidelines is challenging. Defining ethics across contexts requires ongoing dialogue among stakeholders.
	Balancing trade-offs for tangible human-centeredness	AI systems balance objectives like accuracy, efficiency, and fairness, but serving human needs often means trade-offs. Protecting privacy may reduce accuracy. Achieving balance requires careful consideration, posing practical challenges.
	Human-centeredness not an all-encompassing totem	Human-centric AI design is valued but not fully integrated into mainstream development, often overshadowed by technical and market concerns. Human well-being may conflict with philosophical discussions about the environment and other living beings.
(d) Sustainable development goals	Governance framework	The SDGs provide a basis for collaborative mechanisms to ensure responsible and equitable AI use in advancing these goals.
	Business and investments	Private investments, especially from tech giants, significantly shape AI's impact on SDGs. Balancing profit and social-environmental goals is crucial to avoid unintended harm and prioritise the public interest.
	Local communities	Community engagement and empowerment are vital to prevent biases, discrimination, and marginalisation (refer to "Citizen participation").
	Employment	AI's impact on the workforce requires preparation for the future of work, lifelong learning, and inclusive economic opportunities to mitigate negative effects and promote sustainable employment.
	Assessment of sustainable impacts	Assessing AI's impact on SDGs requires robust evaluation frameworks with metrics to measure efficacy, fairness, and sustainability, ensuring benefits outweigh risks.
(e) Citizen participation	Public consultation framework	Establish a structured public consultation framework to elicit citizens' input on SAI risk management and tailored impact assessment.
	Citizen advisory panels	Form citizen advisory panels comprising individuals from diverse backgrounds, including experts, activists, and representatives from marginalized communities.

	Independent audits	Establish an independent auditing and impact assessment mechanism to evaluate SAI systems and algorithms.
	Ethical and safety "checkpoints"	Institute mandatory ethical and safety checkpoints throughout the development and deployment of SAI systems.
	Stress testing and red teaming	Conduct stress tests and red teaming exercises to evaluate the robustness and potential vulnerabilities of SAI systems.

4. Conclusions and future work

In this paper we considered an emerging form of AI, called Symbiotic AI, that is able to engage in a symbiotic relationship with humans. Human-AI symbiosis poses not only several technological challenges but also many philosophical questions. Here we addressed the latter.

In Section 2 we highlighted the main issues concerning the possible foundation of SAI. Crucial philosophical questions were raised about the possibility of thinking about a form of AI capable of entering symbiosis with humans, starting from the specificity that such a technology should exhibit compared to more traditional AI. In particular, in Section 2.1 we questioned the application of biological categories – such as symbiosis – to artefacts. While robotics and AI challenge the traditional distinction between living and non-living based on the capacity of organisms to move autonomously, self-develop and reproduce, there remain some differences that currently appear insurmountable. Also, in Section 2.2, we offered a scan into three patterns of SAI based on as many definitions of intelligent life: the kind of SAI we can conceive of strictly depends on how we think about the continuity between organisation and intelligence. The less continuist our position, the less chance we have of putting down a foundation of SAI in a strong sense, namely from a biological or ontological point of view.

In Section 3 we emphasised the need to redefine our understanding of symbiosis with AI systems to bridge ethics and human-centred computing. Symbiosis, we argue, is more than a static state; it represents a harmonious coexistence between humans and AI, naturally emerging from interactive experiences. We stress the importance of innovative conceptual frameworks that inform strategic policy alternatives. By fostering deeply ingrained symbiosis guided by ethics, we can envision a future where AI enhances the human experience while respecting our values. This perceived beneficial nature of AI, jointly with the value alignment, are necessary preconditions for making AI technologies ethically acceptable.

For the future, the research will involve jurists in order to address the theme of acceptability not only from the ethical but also from the legal viewpoint. Then, legal and ethical acceptability of SAI will need to go through an *operationalization* process in order to be of practical relevance for the design and implementation of SAI applications. High-level principles will be therefore turned into operational definitions that pave the way to technical solutions, e.g., for (partially) automated compliance testing. Operationalization will be accompanied by appropriate modelling of the SAI application in hand. Notably, *socio-technical systems* (STS) are widely recognized as a valuable approach to complex organisational work design that stresses the interaction between people and technology in workplaces [37]. Also, *multi-agent systems* (MAS) [38] are promising from the practical viewpoint since they enable the simulation of possible scenarios and the experimentation of different operational definitions of the legal and ethical acceptability of SAI in a controlled environment. A starting point might be the MAS prototype presented in [39] for the ethical evaluation and monitoring of dialogue systems. Finally, compliance tests might be reformulated as problems that can be addressed with automated reasoning techniques and/or formal methods. Overall, along the direction already explored in, e.g., [40], logic will play a prominent role in the implementation of the computational solutions for many of the problems in our research on SAI within FAIR.

5. Acknowledgements

This work was partially supported by the project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

6. References

1. Grigsby, S.S. (2018). Artificial Intelligence for Advanced Human-Machine Symbiosis. In: Schmorrow, D., Fidopiastis, C. (eds) Augmented Cognition: Intelligent Technologies. AC 2018. Lecture Notes in Computer Science, vol 10915. Springer, Cham. https://doi.org/10.1007/978-3-319-91470-1_22
2. H. James Wilson and Paul R. Daugherty (2019). Creating the Symbiotic AI Workforce of the Future. MIT Sloan Management Review. https://sloanreview.mit.edu/article/creating-the-symbiotic-ai-workforce-of-the-future/
3. Magdalena Paluch (2019). Laying The Foundation For Symbiosis Between Humans And Machines (2019). https://www.forbes.com/sites/forbestechcouncil/2019/12/18/laying-the-foundation-for-symbiosis-between-humans-and-machines/?sh=3f4ab8682997
4. L. Floridi, Technology's In-Betweeness, Philosophy & Technology volume 26, 2013, 111–115.
5. K. F. Lee, C. Qiufan, AI 2041: Ten Visions For Our Future, Penguin Random House, New York, NY, 2021.
6. J. P. Ponda, Artificial Intelligence (AI) through Symbiosis, Thesis, Georgia Institute of Technology, 2022.
7. C. R. Darwin, Old & useless notes about the moral sense & some metaphysical points, CUL-DAR91.4-55, Edited by John van Wyhe, 1838-1840, URL: http://darwin-online.org.uk/content/frameset?pageseq=1&itemID=CUL-DAR91.4-55&viewtype=text.
8. Aristotle, De Anima, Clarendon Press-Oxford University Press, Oxford, 2016.
9. H. R. Maturana, F. J. Varela, Autopoiesis and Cognition. The Realization of the Living, Reidel Publishing Company, Dordrecht, 1980.
10. E. Mayr, The Growth of Biological Thought, Harvard University Press-Belknap, Cambridge, MA, 1982.
11. E. Sober, Learning from Functionalism: Prospects for Strong Artificial Life, in: C. Langton, C. Taylor, J.D. Farmer, & S. Rasmussen (Eds.), Artificial Life II, SFI Studies in the Sciences of Complexity, Proc. Vol. X, Addison-Wesley, Redwood City, CA, 1991, pp. 749–765, DOI:10.1017/CBO9780511730191.021.
12. D. Lévy, Are Robots Alive?, in: A. D. Cheok, E. Y. Zhang, Human–Robot Intimate Relationships, A. D. Cheok and E. Y. Zhang, Human–Robot Intimate Relationships, Human–Computer Interaction Series, Springer Nature Switzerland AG, 2019, pp. 155–188, DOI: https://doi.org/10.1007/978-3-319-94730-3_8.
13. D. E. Koshland, The seven pillars of life, Science, 295(5563) (2002) 2215–2216. DOI: 10.1126/science.1068489.
14. F. M. De Collibus, Are Large Language Models "alive"?, 2023. URL: https://philpapers.org/archive/DECALL.pdf.
15. R. Manzotti, S. Rossi, IO & IA, Rubbettino, Soveria Mannelli, 2023.
16. T. Starner, Using Wearable Devices to Teach Computers, 2017. URL: https://www.youtube.com/watch?v=hi9RYBaPVPI.
17. P. Kotipalli, Symbiotic Artificial Intelligence, 2019. URL: https://p13i.io/posts/2019/06/symbiotic-ai/.
18. R. Saracco, R. Madhavan, S. Mason Dambrot, D. de Kerchove, T. Coughlin, Symbiotic Autonomous Systems. A FDC Initiative, White Paper, Institute of Electrical and Electronics Engineers, 2017.

19. S. Mason Dambrot, D. de Kerchove, F. Flammini, W. Kinsner, L. MacDonald Glenn, R. Saracco, Symbiotic Autonomous Systems. A FDC Initiative, White Paper II, Institute of Electrical and Electronics Engineers, 2018.
20. S. Boschert, T. Coughlin, M. Ferraris, F. Flammini, J. Gonzalez Florido, A. Cadenas Gonzalez, P. Henz, D. de Kerckhove, R. Rosen, R. Saracco, A. Singh, A. Vitillo, M. Yousif, Symbiotic Autonomous Systems. A FDC Initiative, White Paper III, Institute of Electrical and Electronics Engineers, 2019.
21. Macphail, E. M., Barlow, H. B., Vertebrate Intelligence: The Null Hypothesis [and Discussion], Philosophical Transactions of the Royal Society B: Biological Sciences, 308(1135) (1985), 37–51, DOI: https://doi.org/10.1098/rstb.1985.0008.
22. L. Floridi, The Ethics of Artificial Intelligence. Oxford University Press, Oxford, 2023.
23. D. C. Dennett, From Bacteria to Bach and Back. The Evolution of Minds, Penguin Books, London, 2018.
24. Edwards, L., & Veale M. (2017). Slave to the algorithm? Why a right to explanation is probably not the remedy you are looking for. 16 Duke Law & Technology Review 18 (2017), Available at SSRN. DOI: https://doi.org/10.2139/ssrn.2972855.
25. Shin, D., & Park, Y. J. (2019), Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior 98, 277-284. DOI: https://doi.org/10.1016/j.chb.2019.04.019.
26. Selbst, A. D., et al. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency. Fat* 19. ACM Press, 59-68. DOI: https://doi.org/10.1145/3287560.3287598.
27. Wong, P.-H. (2019), Democratizing algorithmic fairness. Philosophy & Technology, 33, 225-244. DOI: https://doi.org/10.1007/s13347-019-00355-w.
28. Katell, M., et al. (2020). Toward situated interventions for algorithmic equity: Lessons from the field. Proceedings of the Conference on Fairness, Accountability, and Transparency. Fat* 19. ACM Press, 45-55. DOI: https://doi.org/10.1145/3351095.3372874.
29. Natale, S. (2021). Deceitful Media: Artificial Intelligence and Social Life after the Turing. OUP.
30. Mota-Valtierra, Georgina, Juvenal Rodríguez-Reséndiz, and Gilberto Herrera-Ruiz. (2019). Constructivism-Based Methodology for Teaching Artificial Intelligence Topics Focused on Sustainable Development. Sustainability 11, no. 17: 4642. DOI: https://doi.org/10.3390/su11174642.
31. Feenberg, A. (2017). Technosystem: The Social Life of Reason. Harvard University Press.
32. Ihde, D. (1990). Technology and the lifeworld: From garden to earth. Indiana University Press.
33. Latour, B. (2005). Reassembling the social: an introduction to actor-network-theory. Oxford University Press.
34. Verbeek, P.-P. (2011). Moralizing technology: Understanding and designing the morality of things. University of Chicago Press.
35. Crawford, K., and Ryan, C. (2016). There Is a Blind Spot in AI Research. Nature 538, 311-313. DOI: 10.1038/538311a.
36. Coeckelbergh, M. (2020). AI Ethics. MIT Press.
37. Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering, Interacting with computers, 23(1), 4-17.
38. A. Dorri, S. S. Kanhere and R. Jurdak, "Multi-Agent Systems: A Survey," in IEEE Access, vol. 6, pp. 28573-28593, 2018, doi: 10.1109/ACCESS.2018.2831228.
39. A. Dyoub, S. Costantini, I. Letteri, F. A. Lisi, A logic-based multi-agent system for ethical monitoring and evaluation of dialogues, in: A. Formisano, Y. A. Liu, B. Bogaerts, A. Brik, V. Dahl, C. Dodaro, P. Fodor, G. L. Pozzato, J. Vennekens, N. Zhou (Eds.), Proceedings 37th International Conference on Logic Programming (Technical Communications), ICLP Technical Communications 2021, Porto (virtual event), 20-27th September 2021, volume 345 of EPTCS, 2021, pp. 182–188. URL: https://doi.org/10.4204/EPTCS.345.32. doi:10.4204/EPTCS.345.32.

40. Dyoub, A., Costantini, S. & Lisi, F.A. Learning Domain Ethical Principles from Interactions with Users. Digital Society 1, 28 (2022). https://doi.org/10.1007/s44206-022-00026-y