

Review

From Human–Machine Interaction to Integration: A Review of Paradigms, Principles, and Prospects for a Human-Centric Future

Sten Grahn^{1,2,3}, Tianyu Wang^{1,*}, Xi Vincent Wang¹ and Lihui Wang¹

¹ Department of Production Engineering, KTH Royal Institute of Technology, 10044 Stockholm, Sweden

² RISE Research Institutes of Sweden, 11428 Stockholm, Sweden

³ Department of Engineering Sciences, Mälardalen University, 72123 Västerås, Sweden

* Correspondence: tianyuwa@kth.se

How To Cite: Grahn, S.; Wang, T.; Wang, X.V.; et al. From Human–Machine Interaction to Integration: A Review of Paradigms, Principles, and Prospects for a Human-Centric Future. *Human-Machine Interaction* **2026**, *1*(1), 1

Received: 17 April 2026

Revised: 2 August 2026

Accepted: 12 August 2026

Published: 21 August 2026

Abstract: As emerging technologies increasingly blur the line between humans and machines, human–machine interaction (HMI) is poised to evolve into deeper forms of human–machine integration. This review examines conceptual frameworks and future scenarios for human–machine systems, focusing on human-centric values and agency. Drawing on foundational models of human–automation interaction and recent advances across multiple disciplines, we delineate integration paradigms from external interfaces and wearables to embedded interfaces and societal-scale intelligent environments, and survey research in each domain. We explore potential futures ranging from human–AI competition to collaboration and symbiosis, to hypothetical merging of humans with intelligent machines. We identify key considerations for steering these futures: defining beneficial outcomes for humanity, preserving human agency and rights as machines gain autonomy, ensuring inclusive benefit for those who do not adopt technological enhancements, and securing foresight capacity to guide integration in line with human values. By synthesizing research across HMI, brain–computer interfaces, Industry 5.0, AI ethics, and adaptive governance, we highlight strategies for human-centric interaction and integration frameworks that can inform technological design and policy. We decompose societal-scale integration into three stages ordered by exit cost and indicate which mapped principles are established, transferred from adjacent domains, or newly proposed. For each research community addressed, we state one implication for reporting or regulatory practice.

Keywords: human–machine integration; human-centric design; human–AI collaboration; brain–computer interfaces; Industry 5.0; human agency and adaptive governance

1. Introduction

Since the dawn of civilization, humans have developed tools and machines to enhance productivity and expand capabilities. Throughout this history, a tacit assumption prevailed: humans remain in control of their tools, directing their use and purpose. Even so, achieving truly human-centric technology has often been challenging; industrial advances brought not only prosperity but also harmful side effects, from widening economic inequality to environmental degradation [1,2]. The relationship between humans and their tools has been shaped by a persistent tension: technology amplifies human capabilities, yet its deployment is rarely neutral in its effects on different segments of society. Today, we stand on the threshold of a new era in which machines are not only more intelligent and autonomous, but may also become deeply integrated into human lives and even bodies. Visionaries



Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

like Kurzweil [3,4] have predicted that within a few decades, the boundary between human and machine could blur as we develop direct brain–computer interfaces and merge with our technologies. These projections raise fundamental questions: How can we ensure that human–machine integration amplifies human well-being and agency, rather than undermining them? What guiding principles and policies are needed now to shape a future in which ubiquitous AI and cyborg technologies remain aligned with human-centric values?

The academic study of human–machine interaction has a rich history. Foundational work by Parasuraman, Sheridan, and Wickens [5] established a model of types and levels of human interaction with automation that continues to inform research and system design. Their framework identified ten levels of automation ranging from full human control to full machine autonomy, with each level entailing different implications for human workload, situation awareness, and performance. Endsley [6] subsequently drew on decades of research to articulate lessons learned from human–automation interaction, emphasizing the persistent challenges of maintaining situation awareness as automation increases in complexity and capability. These foundational models, while developed primarily for traditional external interfaces, provide essential conceptual scaffolding for understanding the deeper forms of integration now emerging. At the same time, this scaffolding should not be carried forward uncritically. These models were developed for supervisory control settings in which a human operator, clearly separated from the machine, allocates functions and monitors discrete levels of automation through external displays. Several of their core assumptions (a well-defined human–machine boundary, discrete and stable levels of autonomy, and oversight exercised through conscious attention) become strained at deeper levels of integration: closed-loop brain–computer interfaces adapt bidirectionally in real time, act on neural signals that users may not consciously produce, and blur the very boundary that supervisory-control models presuppose [7–9]. Throughout this review, we therefore treat these classical models as conceptual scaffolding to be critically extended rather than presupposed.

The urgency of revisiting and extending these frameworks is evident from recent developments. As Vaccaro et al. [10] have shown through a systematic review and meta-analysis, human–AI combinations are useful in specific conditions but often underperform either humans or AI alone, which suggests that a new paradigm for human–AI interaction is needed, one that goes beyond usability to address the deeper cognitive, social, and ethical dimensions of living and working alongside intelligent machines. Rahwan et al. [11] have proposed the study of machine behavior as an interdisciplinary scientific field, arguing that as machines increasingly function as autonomous agents in society, understanding their behavior requires insights from computer science, engineering, psychology, and the social sciences. Similarly, Tsvetkova et al. [12] have called for a new sociology of humans and machines that examines how algorithms, robots, and humans interact in networks and organizations. At the policy level, the European Commission’s Industry 5.0 vision [13] and UNESCO’s Recommendation on the Ethics of AI [14] both emphasize that technological development must be guided by humanistic values: sustainability, resilience, and above all, human-centricity. More recently, the International Monetary Fund [15] has documented how AI and new skills are jointly reshaping labor markets worldwide, which adds urgency to aligning technological integration with workforce adaptation strategies. The Stanford AI Index [16] reports that AI adoption has accelerated dramatically, with generative AI in particular transforming the nature of human–machine interaction across virtually every sector of the economy.

Against this backdrop, four gaps in the existing literature motivate the present review. First, the transition from interaction to integration has so far been examined largely within separate communities. Within human–computer interaction, Farooq and Grudin [17] announced a shift from interaction to integration, understood as a partnership in which humans and proactive software act with autonomy, and a subsequent multi-expert research agenda defined human–computer integration primarily as tight bodily coupling, mapping design challenges around bodily agency and bodily ownership [18,19]. This body of work offers rich design insights at the scale of the individual body, but it does not connect bodily integration to the organizational, industrial, and societal scales at which integration is simultaneously unfolding, nor to the governance instruments that those scales demand. Second, the classical human–automation frameworks discussed above predate the integration paradigm by more than two decades, and their adequacy for closed-loop, body-internal technologies has not been systematically examined. Third, ethical guidance remains fragmented: general AI ethics frameworks [14,20,21] were not formulated for technologies that cross the bodily boundary, while the emerging neurorights literature [8,22] has advanced largely at the level of principle, leaving open how such protections map onto specific forms and depths of integration. Fourth, and cutting across all of the above, the literature lacks an integrated framework that links the depth of human–machine coupling to plausible future trajectories and to the design and policy principles required to keep those trajectories human-centric.

To address these gaps, this review articulates the continuum of human–machine interaction paradigms, from basic external interactions to full integration, providing a framework to evaluate how human roles and rights might change as we move along that spectrum. We then explore future scenarios of human–technology relations, ranging

from competition to cooperation to convergence, that policymakers and designers should anticipate. Finally, we discuss key strategies for safeguarding human interests no matter which path technological evolution takes, including preserving human agency, clearly defining the human benefits that technology should create, and ensuring that even those who choose minimal technological augmentation are not left behind. By examining both the technological trajectory and the socio-ethical guardrails required, this review contributes to an emerging discourse on the responsible development of high-integration human-machine systems. The overarching message is one of proactivity: we must consciously guide HMI's evolution now, while its future forms are still in flux, so that increasing integration translates into human empowerment and societal well-being rather than a loss of human autonomy or a fractured society. In contrast to the human-computer integration agenda developed within HCI [17–19], which centres on the design of body-machine fusion, the framework developed here spans both bodily and societal scales of integration, is grounded in the human-automation and industrial human-centricity traditions [5,6,13], and treats governance and policy as design concerns of equal rank with technology itself.

This review follows a narrative synthesis approach, drawing on literature from several intersecting fields: human factors and ergonomics, human-computer interaction, robotics, neurotechnology, AI ethics, science and technology studies, and technology governance. Sources were identified through searches of major databases (Scopus, Web of Science, IEEE Xplore, PubMed) using terms related to human-machine interaction, human-machine integration, brain-computer interfaces, Industry 5.0, AI ethics, and adaptive governance, supplemented by backward and forward citation tracking of key publications. Given the breadth of the topic, we do not claim exhaustive coverage of any single subfield; rather, we aim to synthesize insights across disciplinary boundaries in order to develop an integrated perspective on the trajectory from interaction to integration and its implications for human-centric design and policy. In the typology of Grant and Booth [23], this article is thus a narrative critical review rather than a systematic review: its aim is conceptual integration across heterogeneous fields, not exhaustive aggregation of empirical effect sizes for a narrowly defined question, for which systematic protocols such as PRISMA [24] are designed. Following established guidance on the complementary roles of the two review genres [25], we prioritized interpretive breadth over procedural exhaustiveness. Sources were included when they (i) constituted foundational or widely cited contributions to their fields, (ii) provided empirical evidence directly bearing on human-centric integration, or (iii) represented authoritative policy or governance documents; no meta-analytic synthesis was attempted, because the reviewed literatures differ fundamentally in methods, outcome measures, and units of analysis. Two limitations follow and should calibrate how the claims of this review are read: literature selection was purposive rather than exhaustive, so individual subfields are represented by salient rather than complete evidence bases and the possibility of selection bias cannot be excluded; and the synthesis is interpretive, so the following sections should be read as a conceptual mapping of the field rather than as quantitative estimates of the balance of evidence. A third limitation follows from the cross-level aim itself. Depth within any single subfield is bounded here by design, and a specialist should not expect this review to advance the state of the art inside their own field. Its contribution is cross-level rather than intra-level, and Section 5 says what a reader working at one level should be able to take from it.

2. From Interaction to Integration: Emerging Paradigms of HMI

Human-machine interaction has traditionally meant humans operating external tools or computers via interfaces like keyboards, screens, or control panels. However, the nature of HMI is rapidly transforming as machines become more capable, autonomous, and physically proximate to (or even embedded within) the human body. Recent taxonomic efforts have sought to organize this expanding field: Onnasch and Roesler [26] proposed a multi-dimensional taxonomy for human-robot interaction that accounts for human, robot, interaction, and contextual factors, while others have begun to conceptualize hybrid intelligence systems that combine human and machine capabilities [27]. Building upon these efforts, and drawing on the classic automation levels model of Parasuraman et al. [5], the current literature can be organized into four analytically distinct levels of human-machine integration (Figure 1 and Table 1). Levels 1–3 range from external machines to wearable and body-internal systems, while Level 4 captures integration into intelligent environments and institutions. Across Levels 1–3, deeper bodily coupling promises greater enhancement but also raises new design and ethical challenges; Level 4 introduces structural constraints at environmental and societal scale. Two clarifications frame this taxonomy. The first is terminological. The classical notion of human-machine symbiosis, introduced by Licklider [28] and echoed in contemporary work on symbiotic collaboration [29,30], denotes a mutually beneficial interdependence of humans and machines: it describes a desirable quality of the human-machine relationship. Integration, as used here, is deliberately the broader and normatively neutral term: it describes the structural depth of coupling between humans and machines, which may or may not prove symbiotic in outcome. In this vocabulary, symbiosis is the goal state

that human-centric design pursues within increasingly integrated systems, and it reappears as the collaborative synergy scenario in Section 3. The second clarification concerns the dimension along which the four levels are ordered. They are arranged not by physical proximity but by the decreasing separability of human and machine: by what it costs to suspend, remove, or exit the coupling. Separability has two components. They coincide at the first three levels and come apart at the fourth. Individual separability is the cost to a person of detaching from a particular system, whether that means walking away from a console at Level 1, taking off a device at Level 2, or accepting explantation and the loss of a restored or augmented function at Level 3. Structural separability is the cost of declining an entire class of technology and still being able to hold a job, obtain credit, and use public services. Levels 1–3 are ordered by individual separability, and structural separability barely enters into them: someone who declines an exoskeleton or a neural implant can in most settings still work and still use public services. At Level 4 the relation inverts. A recommender system or a language assistant is trivially separable as an artifact, since the session can simply be closed, yet the class of system it belongs to may be impossible to decline without giving up employment, credit, or access to public administration. Level 4 is therefore not a departure from the exit-cost axis. It is the point at which the binding component changes from the individual to the structural, and the direction of embedding inverts with it: the machine no longer enters the human body, the human enters a machine environment. This needs stating carefully, because the four levels are not a single scalar scale. Levels 1–3 are ordered monotonically in individual exit cost, and Level 4 is ordered monotonically in structural exit cost (Table 2), but the two ranges are not commensurable, and one consequence is worth stating plainly: a Level 4a system is easier for an individual to walk away from than a Level 3 implant. What the numbering tracks is which component binds, not how large the total cost is. We take this to be a feature of the framework and not an awkwardness in it, since the systems that are individually easiest to close are also the ones whose structural cost is most often overlooked. Level 4 does engage a partly different research community, AI governance and science and technology studies rather than bodily and body-proximal integration, and we treat this bridging role as a deliberate feature of the framework; the compound risks that arise when several levels operate at once are examined at the end of this section. Table 1 states the exit condition that characterizes each level. Figure 2 illustrates the interaction mechanisms and human-centric design foci, reporting individual separability for Levels 1–3 and structural separability for Level 4.

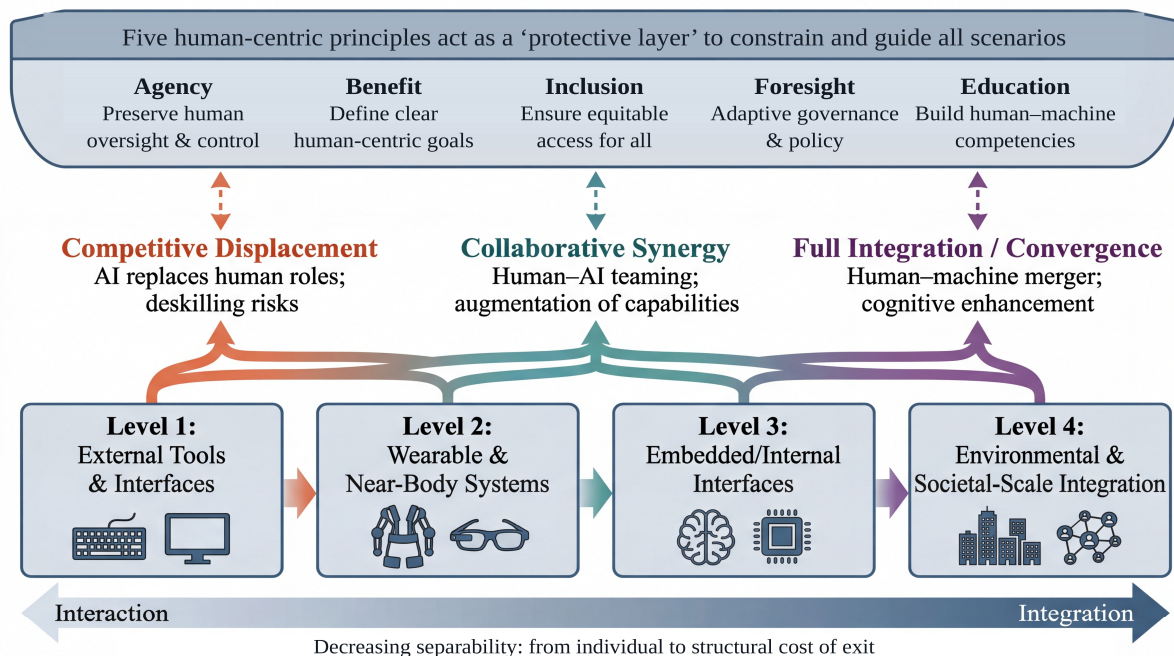


Figure 1. Conceptual framework illustrating the spectrum of human-machine integration levels (**bottom**), three plausible future scenarios (**middle**), and the cross-cutting human-centric principles that should guide technological and policy development (**top**). The horizontal axis is the separability of human and machine, that is, the cost of exiting the coupling. Levels 1–3 are ordered by individual exit cost and Level 4 by structural exit cost; the two are not commensurable, as Section 2 explains.

Table 1. Levels of human–machine integration, with examples, exit conditions, and human-centric design considerations.

Level of Integration	Description and Examples	Separability: What Exit Requires	Human-Centric Considerations
External Tools and Interfaces	Traditional HMI: external machines operated via input devices (e.g., computers with screens/keyboards; robots controlled by console) [5,26].	Ceasing to operate the machine. The coupling ends when the user steps away, and the cost of exit is limited to the loss of the task benefit.	Focus on usability, ergonomic design, and ensuring machine outputs are transparent and interpretable to human operators. Trust calibration and appropriate reliance are critical [31,32].
Wearable and Near-Body Systems	Machines worn on, attached to, or in close proximity to the body without establishing a direct internal or neural interface (e.g., external prosthetics, AR glasses, exoskeletons, smart contact lenses) [33,34].	Removing or detaching the device. The cost is negligible for non-permanent wearables such as AR glasses and exoskeletons, and may be substantial where a body-attached device restores a lost function, as with advanced external prosthetics.	Enhance human capabilities (strength, senses, data access). Key issues: safety and reliability; ethical use of enhancements; equitable access to avoid a “tech-augmented” vs. “non-augmented” divide [35].
Embedded/Internal Interfaces	Technologies integrated within the human body or brain for direct communication (e.g., brain–computer interfaces; neuroprosthetic implants enhancing cognition or senses) [7,36–38].	Explantation, deactivation, or loss of a restored or augmented function. Separation carries surgical, clinical, or functional cost and may be irreversible.	Enable highly intuitive interaction and potential cognitive expansion. Key issues: preservation of mental privacy and autonomy [8,22]; security against misuse; rigorous validation of safety; regulatory standards for human augmentation.
Environmental and Societal-Scale Integration	Humans interfacing with intelligent environments or large-scale automated systems (smart homes/cities, autonomous vehicles, global AI networks) [39,40].	Exit is structural rather than physical. It requires declining an entire class of system and, with it, the work, services, or civic participation that the system mediates. Sub-levels 4a–4c order this cost (Table 2).	Allow human oversight of complex, distributed systems via AI assistance. Key issues: trust and transparency in large-scale AI decisions; maintaining human control in critical infrastructure [41]; safeguarding public interest.

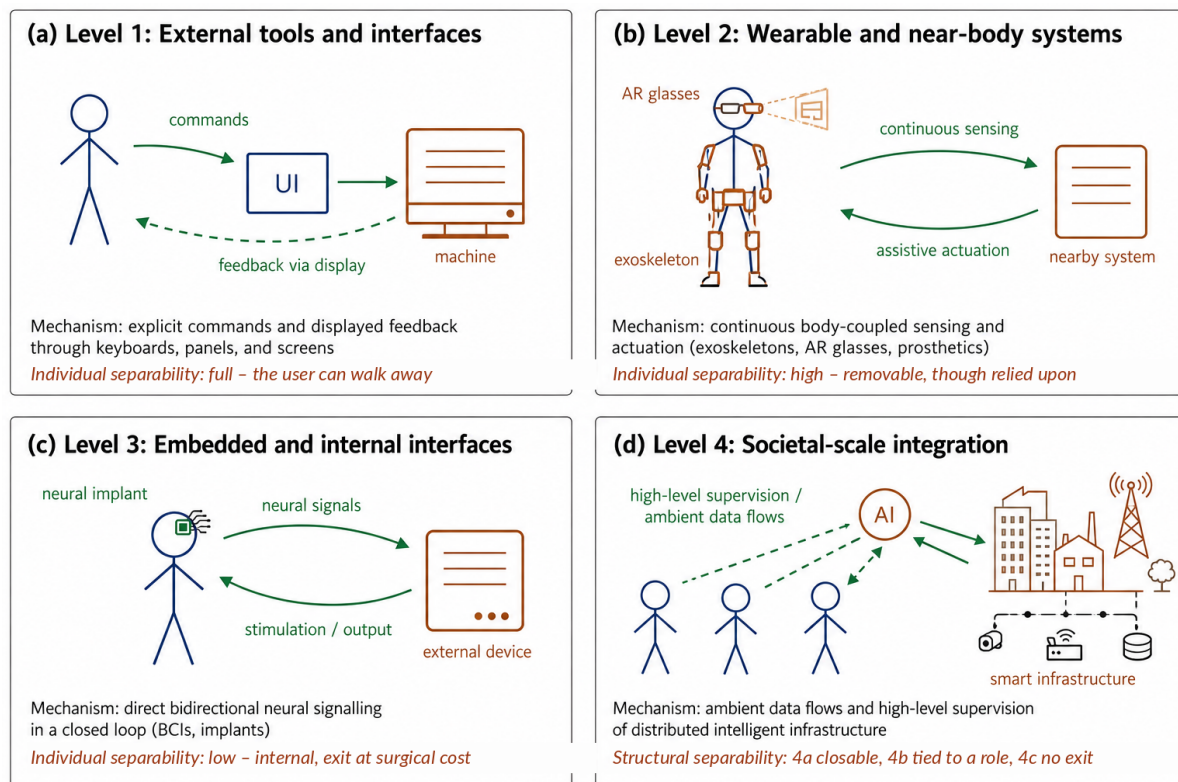


Figure 2. Schematic illustration of the four levels of human–machine integration, showing the characteristic interaction mechanism and the dominant human-centric design considerations at each level: (a) external tools and interfaces; (b) wearable and near-body systems; (c) embedded and internal interfaces; (d) environmental and societal-scale integration. Panels (a–c) report individual separability, the cost to a person of detaching from the system; panel (d) reports structural separability and names its three stages (Table 2).

Integration level 1. At the first level, humans interact with machines through explicit external interfaces. This is the status quo of most workplace and consumer technology: we type on keyboards, click buttons, use touchscreens, or manually operate machinery through control panels. Shneiderman [31] has articulated a

comprehensive vision for human-centric AI at this level, arguing that AI systems should be designed with reliable, safe, and trustworthy interfaces that enhance human performance while maintaining human control. A critical challenge at this level is calibrating trust, the degree to which human operators rely on automated recommendations or outputs. Lee and See [32] provided a seminal analysis of trust in automation, showing that trust is shaped by factors including system reliability, the operator's understanding of how the system works, and the perceived risk of errors. Their work established that both over-trust (leading to complacency and automation bias [42]) and under-trust (leading to disuse of beneficial automation) can degrade system performance. Hancock et al. [43] extended this analysis through a meta-analysis of factors affecting trust in human–robot interaction, finding that robot-related factors (particularly performance and reliability) have the strongest effect on trust, while human-related factors such as personality and expertise play a secondary role. A recent systematic review by Mehrotra et al. [44] extends these findings to AI systems specifically, identifying design interventions, user training strategies, and transparency mechanisms that can support calibrated trust across a range of human–AI interaction contexts. More recently, Dittmar and Sposato [45] have proposed a trust and capability framework for optimizing human–AI decision performance in knowledge management, while Bansal et al. [46] have demonstrated experimentally that the form and content of AI explanations significantly modulate team performance, trust, and reliance behavior in complex decision tasks. Glikson and Woolley [47] synthesize empirical research on human trust in AI, emphasizing that trust is not a static property but a continuously negotiated relationship shaped by representation, tangibility, and task context. These findings carry direct implications for the design of interfaces at every level of integration: systems must be designed not merely to function well, but to encourage appropriate levels of reliance by their human users.

Integration level 2. The second level involves wearables and near-body systems, which are already proliferating across both consumer and industrial domains. This category includes devices that people carry, wear, or attach to the body, as well as those in close proximity, provided they do not establish a direct internal or neural interface. We distinguish between body-attached systems and non-permanent wearables. Body-attached technologies include advanced external prosthetic limbs; future systems may enhance capabilities beyond therapeutic restoration. Non-permanent wearables include AI-enhanced eyeglasses, smart contact lenses with potential future visual-enhancement capabilities, and exoskeleton suits that augment strength and endurance. Occupational exoskeletons illustrate how near-body technology can augment human physical capabilities: research has shown significant potential for reducing physical strain in industrial settings [33], with recent roadmaps charting a path toward large-scale adoption across manufacturing sectors [34]. Romero et al. [35] proposed the concept of the Operator 4.0, a smart, skilled operator who performs cooperative work with machines through a range of augmentation technologies, including physical (exoskeletons), cognitive (intelligent personal assistants), sensory (augmented reality), and virtual (immersive simulation) enhancements. This typology highlights that wearable and near-body systems serve not merely as tools but as extensions of the operator's capabilities along multiple dimensions. Recent advances in human behavior prediction are further enabling smooth and safe human–robot collaboration in industrial settings, moving toward the Industry 5.0 vision of intuitive, anticipatory partnerships between workers and machines [30]. A comprehensive bibliometric analysis by Haghighi et al. [48] confirms that safety has emerged as the core research focus in collaborative robotics, while Shah et al. [49] document how AI-enhanced cobots are moving toward safer, smarter, and more human-centric industrial automation through generative models for hazard synthesis, predictive reasoning for human motion, and adaptive control for dynamic risk reduction. Patalas-Maliszewska et al. [50] further examine how safety improvements in human–robot collaboration contribute to sustainable production goals within the Industry 5.0 paradigm. Most comprehensively, Wang et al. [51] provide an in-depth review of human-centric assembly in smart factories, addressing principal challenges of human centricity, sustainability, and resilience, and proposing solutions that combine augmented robots, cognitive systems, mixed reality, collaborative intelligence, and brain–robotic controls to reduce both physical and intellectual burdens on human operators.

At this level, technology remains external or removable, but it begins to function as a natural extension of the user's body or senses. From a human-centric perspective, key concerns include device safety, interoperability with the human body, and ensuring that those who rely on such enhancements remain in control of them. Ethical questions also arise with particular urgency: how do we equitably manage access to augmentative technologies so that society does not fracture into the augmented and the unenhanced? The risk of creating new forms of social stratification based on technological enhancement mirrors historical patterns identified in the economics of technological change [1,52]. Shin [53] has recently synthesized the literature on human-centric AI and digital well-being, arguing that the design of AI systems must attend not only to functional performance but also to users' psychological autonomy, sense of meaning, and long-term flourishing. Ensuring that human dignity and

opportunities are preserved for people who cannot or choose not to use these devices is an important policy challenge that requires proactive attention well before such technologies become widespread.

Integration level 3. At the third level, embedded or internal interfaces move interaction inside the human body. The defining boundary between Levels 2 and 3 is the interface: implanted, internal, or direct neural interfaces are Level 3 regardless of device size, whereas externally worn or attached systems without such interfaces remain Level 2. Pioneering efforts in brain–computer interfaces (BCIs) demonstrate early steps toward direct communication between the human brain and machines. Musk and colleagues at Neuralink [36] have described an integrated BCI platform with thousands of channels. Neuralink’s ongoing PRIME early-feasibility study is evaluating the N1 implant in participants with tetraparesis or tetraplegia, and sponsor-reported early use has included thought-based control of external devices [37,38]. Saha et al. [9] provide a comprehensive review of progress and remaining challenges in BCI research, documenting how the field has progressed from laboratory proof-of-concept to clinically meaningful applications across motor, sensory, and cognitive domains. Earlier foundational work by Lebedev and Nicolelis [7] reviewed the physiological and engineering principles underlying BCIs and charted a path from basic science to neuroprostheses and neurorehabilitation, distinguishing between non-invasive approaches (such as electroencephalography-based interfaces) and invasive ones (such as intracortical electrode arrays) and assessing their respective trade-offs in terms of signal resolution, surgical risk, and long-term stability. Such closed-loop systems also expose the limits of the classical human–automation models introduced in Section 1: in a bidirectional neuroadaptive interface there is no discrete level of automation to select, no external display through which oversight is exercised, and no stable boundary at which the human ends and the machine begins, which is precisely why integration-specific ethical and design frameworks are required [8,9,22].

In the ideal envisioned by integration enthusiasts, one might someday control devices by thought alone, with digital tools acting as true cognitive prostheses. Achieving this kind of frictionless interface could eliminate the conscious effort of operating technology; instead of typing or speaking, a person might simply imagine an action and the implanted system would execute it. While this promises unprecedented levels of convenience and ability (for example, restoring mobility to the paralyzed, or even enhancing memory and computation for healthy users), it raises profound risks. Preserving human agency and mental privacy becomes paramount: scholars such as Yuste et al. [8] have identified four ethical priorities for neurotechnologies (privacy, identity, agency, and augmentation), while Ienca and Andorno [22] have proposed new human rights for the age of neuroscience, including rights to cognitive liberty, mental privacy, mental integrity, and psychological continuity. If machines are tied into our neural processes, strict safeguards are needed so that they augment rather than override human intentions. There are also questions of personal identity (how do these integrations affect one’s sense of self?), liability (if an implanted AI influences someone’s decision, who is responsible?), and reversibility (can users meaningfully choose to disconnect?). Regulatory frameworks will need to address these issues by treating BCIs and similar technologies with rigorous oversight, potentially requiring clinical trials, post-market surveillance, and usage guidelines that prioritize user consent and control. The rapid commercialization of BCIs has further intensified these concerns: Boonstra [54] warns that the premature commercialization of neurotechnologies risks outpacing neuroscientific understanding, while Ienca and Haselager [55] have called for dedicated neurosecurity and legal regimes for neural data processing that go beyond existing data protection frameworks.

Integration level 4. The fourth level extends beyond individual bodies to environmental and societal-scale integration. Here humans do not so much operate a machine as inhabit one: everyday activity is mediated by large intelligent systems made up of many interconnected devices, models, and autonomous agents. Present-day instances include consumer assistants and recommender systems, algorithmically managed workplaces and platforms, and smart-city infrastructure in which traffic control, public transit, power grids, and administrative processes are guided by AI [39,40]. OECD analyses [56] document how AI-driven urban systems are already scaling from individual cities to networked metropolitan regions. Zheng et al. [57] show how industrial knowledge graphs and multi-agent decision-making enable self-organizing cognitive manufacturing networks, which is the kind of distributed intelligence that characterizes this level. At the urban scale, Yigitcanlar et al. [58] find that effective AI-driven urban systems require governance frameworks that actively cultivate public trust and participation. Prospective developments, such as supervising swarms of robots or software agents through high-level commands, or monitoring resource systems at planetary scale, extend this level further, but they are not what defines it. Elements of Level 4, especially Level 4a, are already pervasive.

Because Level 4 spans systems as different as a language assistant and a national benefits administration, it needs the internal differentiation that Levels 2 and 3 already have, where permanence and invasiveness respectively subdivide the level. The separability axis supplies one. Order Level 4 systems by structural exit cost, that is, by what a person must give up to live without the class of system rather than without the particular instance, and three empirically distinguishable stages emerge, summarized in Table 2.

Level 4a, discretionary ambient AI. The system mediates activity, but a person can decline it and still hold a job, transact, and use public services, usually at some cost in convenience or performance. Consumer language assistants, recommender systems, and voluntary smart-home automation belong here. Exit remains available to the individual, and a non-algorithmic pathway to the underlying function still exists.

Level 4b, institutionally mandated AI. The system is interposed between a person and an institution that controls access to an economic or social role: algorithmic hiring and performance management, work allocation on labor platforms, credit and insurance scoring, algorithmic scheduling. The individual cannot decline the system without declining the role, so exit cost has migrated from the artifact to the institution. Aloisi and De Stefano [59] document how algorithmic management restructures the employment relationship in precisely this way, and the European Data Protection Supervisor [60] responds to it by recommending that, where operators rely heavily on system output, oversight duties be discharged as though the system were operating autonomously. This makes 4b a category defined partly by the employment or contractual relation and not by the technology alone. That is the point of the stage: the exit cost has left the artifact, so a taxonomy that classified only artifacts would have nowhere to put it.

Level 4c, infrastructural AI. The system administers a function for which no non-algorithmic pathway remains within the polity: traffic and energy management, identity and benefits administration, emergency dispatch. Exit is available neither individually nor institutionally, but only by leaving the jurisdiction. This is the limit of the separability axis, and it is where the classical assumption of a supervisory operator standing outside the system is least tenable [6,60,61].

Table 2. Decomposition of Level 4 into three stages ordered by structural exit cost, with the marker that distinguishes each stage from the next.

Sub-Level	Exit Condition (Structural Separability)	Representative Systems	Present Status and Transition Marker
4a. Discretionary ambient AI	The individual can decline the system and retain access to work, services, and civic life; the cost of exit is convenience or performance.	Consumer language assistants; recommender systems; voluntary smart-home automation [16,62].	Pervasive today. The transition to 4b occurs when an institution begins to condition access to a role on use of the system.
4b. Institutionally mandated AI	The individual cannot decline the system without forgoing an economic or social role; exit cost lies with the institution rather than the artifact.	Algorithmic hiring and performance management; platform work allocation; credit and insurance scoring; algorithmic scheduling [59,60].	Established in platform work and spreading into standard employment. The transition to 4c occurs when the non-algorithmic pathway for the function is discontinued.
4c. Infrastructural AI	No non-algorithmic pathway remains within the polity; exit requires leaving the jurisdiction.	Traffic and energy management; identity and benefits administration; emergency dispatch [39,40,56].	Emerging unevenly across jurisdictions. Marks the limit of the separability axis, where oversight must be structural rather than procedural [61,63].

This decomposition also answers a question that a single-bucket Level 4 cannot: how a more deeply integrated future would differ from a present in which AI is already pervasive. The difference lies in where essential dependencies sit across 4a, 4b, and 4c, not in capability or ubiquity. Contemporary society sits predominantly at 4a with a rapidly expanding 4b: large language models are used at scale in professional work [16,62], and algorithmic management is established in platform employment and spreading into standard employment [59], while most essential public functions still retain a non-algorithmic pathway. Read together, the literature reviewed here points less to the arrival of a new technology than to the withdrawal of alternatives, the point at which the non-algorithmic route to employment, credit, clinical triage, or public administration is discontinued because too few people use it to justify its cost. The trajectory then becomes something one can observe instead of only project. Three indicators track it: the proportion of essential services for which a non-algorithmic pathway remains available and usable; whether declining a system remains an individual choice or is institutionally penalized; and whether oversight of 4b and 4c systems is exercised by operators with a genuine capacity to override [60,63]. The obligation that follows is to keep alternatives funded and usable. The primary responsibility for maintaining such alternatives lies with public authorities and deploying institutions [14], and it is not an argument for slowing adoption.

The critical human-centric question is how to maintain effective human oversight and input in these massive, automated systems. Taeihagh [41] has examined governance mechanisms for AI in such contexts, emphasizing that adaptive and participatory approaches are needed to keep pace with rapid technological change. Design considerations include creating intuitive control interfaces for supervisors and ensuring that automated systems defer to human-defined limits (for example, a smart grid AI might be given rules that prioritize safety and equity, such as preventing life-critical services from ever being shut off for payment issues). Additionally, careful thought must be given to cybersecurity and resilience: as humans become more dependent on ubiquitous AI-driven

infrastructure, the vulnerability to system failures or malicious attacks could have widespread human consequences. The challenge of maintaining situation awareness across large, distributed systems echoes the concerns Endsley [6] raised for individual human–automation interaction, but amplified to a societal scale. The European Data Protection Supervisor [60] has recently addressed this challenge directly, arguing that where human operators rely heavily on system recommendations, deployers should apply oversight measures as though the system were operating autonomously, effectively inverting the default assumption from supervised to unsupervised operation. Wood [61] examines whether AI-enabled autonomous systems function as extensions or replacements of human will and control, concluding that maintaining meaningful human agency requires structural design constraints rather than merely procedural safeguards.

Finally, the four levels are analytically distinct but empirically concurrent. A single production worker may already operate machinery through conventional control panels (Level 1), wear an exoskeleton and augmented-reality glasses (Level 2), carry a medical implant (Level 3), and work inside a facility whose scheduling, logistics, and safety systems are algorithmically orchestrated (Level 4). Such multi-level configurations generate compound challenges that a level-by-level analysis misses. Ethically, consent becomes layered and partially illusory: a worker may consent to each device individually yet have no practical ability to refuse the data flows that arise when, for example, exoskeleton telemetry feeds facility-level productivity analytics. In governance terms, responsibility fragments across device manufacturers, employers, software vendors, and infrastructure operators, so that accountability for harms arising from level interactions (for instance, a Level 4 scheduling system driving a wearer beyond safe exoskeleton-assisted workloads) falls between regulatory regimes that address devices, workplaces, and algorithms separately. In control terms, higher levels constrain lower ones: societal-scale systems shape the options available to individuals faster than individuals can adapt, amplifying lock-in and eroding the practical value of device-level override rights. These compound risks reinforce the need for cross-level impact assessment and for principles that apply across levels rather than to any single one, which we develop in Section 4; the cross-level research gap is revisited in Section 8.

3. Future Scenarios: Steering Human–Machine Relations Toward Collaborative Synergy

The trajectory of human–machine relationships could branch into multiple future scenarios, often characterized along a spectrum from competition to cooperation to full integration (see Figure 3). Proactively considering these scenarios is essential for guiding us toward the most desirable outcomes [63]. As Raisch and Krakowski [64] have articulated, organizations and societies face a fundamental automation–augmentation paradox: the same technologies that can replace human capabilities can also be designed to enhance them, and the choice between these paths is often determined more by organizational strategy and policy choices than by the inherent properties of the technology itself. We describe three broad scenarios:

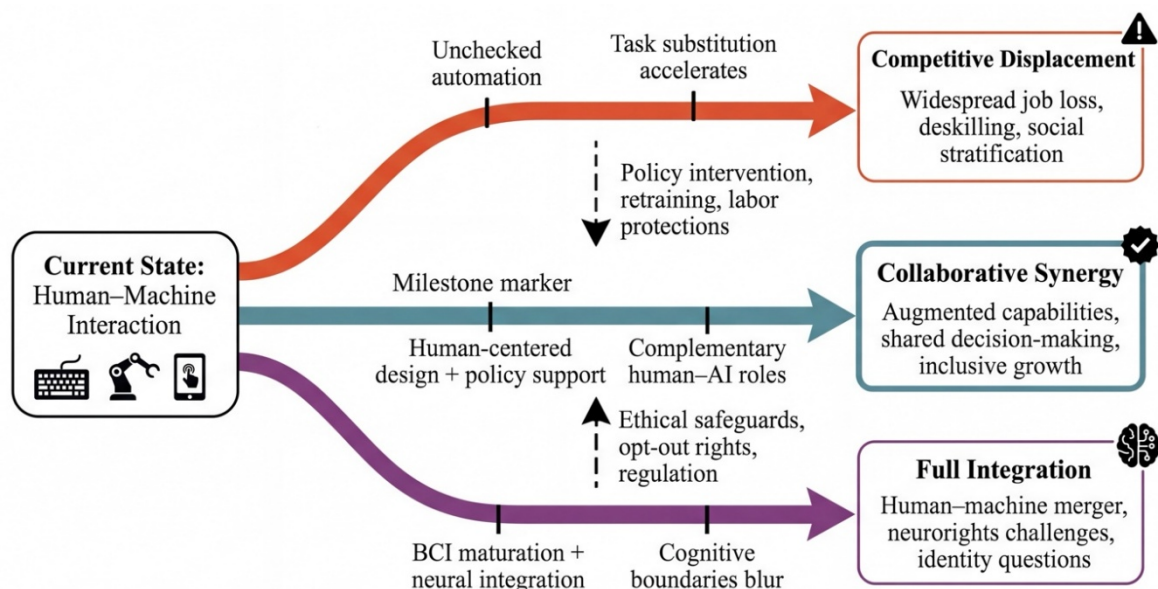


Figure 3. Three future scenarios for human–machine relations. Solid arrows represent developmental trajectories from the current state of human–machine interaction; dashed arrows indicate how policy and design interventions can redirect outcomes toward the collaborative synergy pathway.

Competitive Displacement. In a dystopian scenario, autonomous entities, ranging from advanced AI and robotics to fully autonomous corporations granted legal personhood, effectively compete against humans, displacing human roles and decision-making and even challenging human rights and human-centric governance. As AI advances make it feasible to operate organizations with minimal or no human involvement, the legal personhood already afforded to corporations could evolve into a direct challenge to individual human rights and agency. Such entities might surpass human capabilities in so many domains that humans are sidelined economically, socially, and even cognitively. Early analyses by Frey and Osborne [65] estimated that 47% of US employment was susceptible to automation, sparking widespread concern. More nuanced research by Acemoglu and Restrepo [66] has since shown that automation both displaces existing tasks and creates new ones, though the net effect depends critically on policy choices and the pace of technological change. Autor [67] offered a complementary perspective, arguing that automation tends to substitute for routine cognitive and manual tasks while complementing non-routine analytical and interpersonal work, but cautioning that this complementarity is not guaranteed and depends on institutional arrangements. More recent empirical work by Autor et al. [68] has traced the historical origins and content of new work from 1940 to 2018, demonstrating that while automation consistently eliminates some occupations, it simultaneously generates entirely new categories of employment, a dynamic that AI is now accelerating at unprecedented speed. Empirical evidence from customer-support settings shows that generative AI can substantially augment worker productivity, with particularly large gains among less experienced workers [62]. This indicates that institutional choices about labor regulation, collective bargaining, and social protection are at least as important as technology itself in determining employment outcomes. Brynjolfsson and McAfee [52] further warned of a potential “great decoupling” between productivity growth and employment, in which technology-driven productivity gains accrue primarily to capital owners rather than workers. Aloisi and De Stefano [59] show how algorithmic management and platform work reshape employment relationships and distribute workplace risks. More recent labor-market evidence [15] indicates uneven exposure, with AI-vulnerable and entry-level roles facing particular pressure even while aggregate effects remain uncertain. Acemoglu [69] similarly concludes that the productivity gains from AI have thus far been concentrated in a narrow set of applications, and that broader economic benefits depend critically on whether AI is deployed to augment or replace human labor.

Bostrom [70] has explored extreme scenarios in which superintelligent AI could pose existential risks to humanity if not properly aligned with human values. Even short of such extremes, the competitive displacement scenario carries significant risks of social destabilization: unemployment, social stratification, erosion of purpose and meaning in work [2], and ethical crises in which AI decisions override human preferences. A particularly insidious dynamic is the gradual deskilling of human operators as they become increasingly dependent on automated systems, a concern documented in the human factors literature on automation complacency and bias [42]. Such an outcome might emerge if society fails to put human-centric guardrails on AI development, allowing profit or efficiency motives to dominate unchecked. Recent data from the International Monetary Fund [15] and the World Economic Forum [71] confirm that these dynamics are already materializing: the WEF estimates that while 92 million jobs may be eliminated by 2030, 170 million new roles could emerge, though this net gain is by no means guaranteed without deliberate policy intervention. Employment levels in AI-vulnerable occupations are measurably lower in regions with high demand for AI skills, and entry-level positions show especially high exposure to generative AI, creating particular challenges for young workers entering the labor market. Avoiding this path requires early intervention: setting firm policies that prohibit fully autonomous systems in roles that could marginalize or harm humans, investing in education and retraining to prepare workforces for new roles, and ensuring continuous opportunities for meaningful human contribution.

Collaborative Synergy. A more optimistic scenario is one of human–machine cooperation, where AI and automation are deployed as partners or assistants to humans. This scenario represents the augmentation side of the automation–augmentation paradox [64]: rather than replacing humans, technology is deliberately designed to enhance their capabilities. Dellermann et al. [27] have described this as hybrid intelligence: systems that combine human and machine capabilities to achieve results superior to either alone. The concept rests on the observation that human and machine intelligence have complementary strengths: machines excel at processing large volumes of data, maintaining consistency, and performing rapid computation, while humans bring contextual understanding, ethical judgment, creative reasoning, and the ability to handle ambiguity [29]. O’Neill et al. [72] provide a systematic review of empirical research on human–autonomy teaming, showing that effective collaboration depends on factors such as trust calibration, shared mental models, and appropriate task allocation between humans and machines. Jarrahi [29] further argues that human–AI symbiosis in organizational decision making requires understanding and drawing on the distinct strengths of each partner. Recent empirical work by Hemmer et al. [73] has formalized this notion of complementarity, showing that while the theoretical potential for

human–AI teams to outperform either partner alone is well established, achieving such complementary performance in practice remains rare and depends on careful task design and mutual calibration. Seeber et al. [74] outline a research agenda on machines as teammates, arguing that as AI systems take on collaborative roles they must increasingly be treated not merely as instruments but as partners in team processes. However, recent analyses caution that current human–AI collaboration remains far less collaborative than the rhetoric suggests, with most deployed systems still following asymmetric patterns in which the AI proposes and the human merely approves, rather than genuine bidirectional collaboration [73,75]. This asymmetry also reflects an acute educational gap: without training in critical evaluation, effective prompting, and collaborative reasoning with AI, the majority of users risk becoming passive consumers of machine outputs rather than genuine intellectual partners. Addressing this gap through reformed educational curricula and workplace training programs is essential if the collaborative synergy scenario is to be realized at scale. The rapid proliferation of generative AI and large language models has added new dimensions to this challenge: Susarla et al. [76] argue that generative AI represents a qualitatively different form of technology whose societal impact demands new research paradigms and responsible scholarly practices in information systems, while Bender et al. [77] identify persistent risks in scaling large language models within socio-technical systems, including issues of reliability, contextual understanding, and the risk of over-reliance. Bai et al. [78] provide experimental evidence that LLM-generated messages can persuade humans on policy issues at rates comparable to human-authored persuasion, raising important questions about the autonomy of human judgment in AI-mediated environments.

In this vision, routine tasks are automated, freeing people for more creative, strategic, and interpersonal work. Human labor shifts toward complementary roles that emphasize empathy, complex problem-solving, supervision, and innovation, aligning with a common characterization of the ideal division of labor: AI does the work, humans do the thinking (or the learning, creating, and caring). This scenario underpins many calls for democratizing AI and ensuring meaningful human control. It reflects the goals of Industry 5.0, which envisions industries using advanced technology to empower workers and improve human well-being rather than simply cutting costs [13,79,80]. Wang [81] has argued for a futuristic perspective on human-centric assembly that places the worker at the center of manufacturing design, a vision further substantiated by a comprehensive review of human-centric assembly in smart factories that demonstrates how augmented robots, cognitive systems, and collaborative intelligence can jointly address the challenges of human centricity, sustainability, and resilience [51]. The Operator 4.0 concept [35] provides a concrete typology for how different augmentation technologies can support workers. Leng et al. [80] offer a comprehensive prospect and retrospect of Industry 5.0, identifying key research challenges including the development of human-centric collaborative manufacturing systems and the need for new metrics that capture human well-being alongside traditional productivity measures. Nahavandi [82] articulates Industry 5.0 as a fundamentally human-centric solution that places worker well-being at the core of manufacturing transformation, while Xu et al. [83] have explored how embodied AI and large language models can be integrated into cyber-physical-social systems to realize human-centric smart manufacturing. Longo et al. [84] articulate a value-oriented and ethical engineering perspective for designing human-centric factories of the future in Industry 5.0, while Gaffinet et al. [85] provide a systematic literature review and concept disambiguation of Human Digital Twins (HDTs), which serve as digital replicas mirroring human characteristics to enhance human-system integration in manufacturing. Miller and Spatz [86] develop a unified view of the Human Digital Twin, articulating how individual-level models integrate sensing, prediction, and adaptive system configuration to enhance human–system collaboration. Maddikunta et al. [87] survey the enabling technologies and potential applications of Industry 5.0, illustrating how smart manufacturing is evolving to integrate human-centric design principles with advanced digital technologies. Achieving this scenario requires deliberate socio-technical design: developing safe and intuitive robots and systems for collaborative applications, reorganizing workplaces for human–AI co-creation, and adopting policies (like shorter work weeks, retraining programs, and social safety nets) so that automation’s benefits translate into improved quality of life for all. It also requires building a culture of trust, something that transparent AI, strong data privacy, and well-communicated governance frameworks can help achieve [32,88]. Microsoft’s New Future of Work Report [89] provides evidence that when workers’ expertise and perspectives inform the design and deployment of workplace AI, organizations achieve more sustainable improvements in both productivity and well-being, reinforcing the case for participatory approaches to human–AI system design.

Full Integration (Human–Machine Merger). The most radical scenario involves an eventual merging of humans and machines. Here, the distinction between human and artificial intelligence might fade, as technologies like BCIs, genetic engineering, and nanotechnology enable people to integrate AI directly into their bodies and minds. Proponents of this scenario see it as the path to potentially transcending human limitations (eliminating disease and aging, vastly expanding intelligence, and perhaps even achieving forms of collective consciousness or

brain-to-brain communication). Sorgner [90] has explored the philosophical and ethical dimensions of this trajectory, arguing that we have always been cyborgs in some sense (given our dependence on language, writing, and cultural tools) and that the key question is not whether we will integrate with technology but how we will govern that integration. This perspective challenges essentialist views of human nature and forces us to confront questions about what aspects of humanity, if any, should be considered inviolable.

A fully integrated world raises profound questions about identity, rights, and social organization. Would augmented individuals enjoy unfair advantages over others? How do we define individual identity, responsibility, or culpability when cognition can be distributed across human–AI networks? Even in a scenario of deep integration, a human-centric stance demands that core human values (such as autonomy, consent, equity, and well-being) remain non-negotiable constraints on technological design [20,91]. Regulatory bodies might require that brain implants cannot manipulate a person’s behavior against their will, and that people retain the right to disconnect from integrated systems at any time. Society may also decide on limits to integration, such as banning certain high-risk enhancements or autonomous AI agents that could erode human responsibility or privacy. If this scenario moves from science fiction to reality, continuous ethical and public deliberation will be needed on questions such as: Do humans need protected cognitive spaces free from any machine intervention? And how do we ensure those who remain unenhanced have equal rights and opportunities? Matthews et al. [92] have recently applied stakeholder theory to the analysis of AI and algorithmic systems, providing a framework for systematically identifying all affected parties and their interests when deploying integrative technologies. UNICEF’s report on neurotechnology and children’s rights [93] further highlights that vulnerable populations, including minors, require additional protections as neural interfaces move toward consumer markets.

These scenarios are not mutually exclusive, and the future may include elements of all three. Different domains may experience different dynamics; for example, some jobs might be fully automated (displacing humans), while others become collaborative, and some individuals opt for deep integration via implants. As Raisch and Krakowski [64] emphasize, the same technology may serve substitution or augmentation purposes depending on how it is deployed. The role of policy is to steer these outcomes toward the collaborative and human-enhancing end of the spectrum and away from the competitive, human-displacing end [21]. This requires foresight to recognize emerging trends and set boundaries early on undesirable practices, while encouraging technologies and business models that reinforce partnership between humans and machines. Methodologically, the three scenarios are analytic constructs derived from the literature reviewed above rather than free-form speculation. Their common axis is the direction in which the automation–augmentation paradox [64] is resolved: technologies deployed primarily to substitute for human capabilities point toward competitive displacement, technologies deployed to complement them point toward collaborative synergy, and technologies that dissolve the human–machine boundary altogether point toward full integration. Each scenario is likewise anchored in a distinct empirical literature: respectively, the labor economics of automation [52,65–67], research on human–autonomy teaming and hybrid intelligence [27,29,72], and the neurotechnology and transhumanism literatures [7,36,90]. The scenarios also differ in the observable indicators by which their emergence can be tracked. Sustained divergence between productivity and median labor income, and displacement without task reinstatement, would evidence competitive displacement [52,66]; growing complementary human–AI team performance and stable or rising labor shares would evidence collaborative synergy [73]; and mainstream adoption of invasive augmentation by healthy users would evidence movement toward full integration [54]. Historical precedent counsels against treating any one trajectory as inevitable: successive waves of mechanization and computerization produced displacement and complementarity simultaneously, with institutional choices determining the balance [67,68].

4. Key Principles for Human-Centric Integration

The preceding review of integration levels and future scenarios reveals recurring tensions: between human capability and machine efficiency, between individual enhancement and social equity, and between the pace of innovation and the capacity of governance to keep up. These tensions point toward a set of cross-cutting principles that recur across the literature, regardless of which scenario is considered or which level of integration is at stake. Synthesizing the ethical frameworks that have emerged across AI ethics [20,21], bioethics [8,22], and technology governance [41,94], the following five principles represent an emerging consensus on what human-centric integration demands (see Table 3):

Table 3. Summary of five cross-cutting principles for human-centric integration, with core objectives and key strategies.

Principle	Core Objective	Key Strategies and Mechanisms
Preserve Human Agency and Rights	Ensure humans retain meaningful control and accountable authority in consequential domains.	Human override capabilities; hard constraints on AI behavior; opt-out rights; accountability frameworks that prevent autonomy gaps [6,63,94,95].
Clearly Define Human Benefit	Ground technology development in explicit, value-based human welfare goals.	Interdisciplinary value definition; beneficence and explicability metrics; integration of ecological sustainability [20,96].
Ensure Inclusive Technology and Social Cohesion	Broadly share the benefits of integration and protect those who remain unaugmented.	Accessible service alternatives; anti-discrimination protections; publicly funded augmentation programs; rights for non-adopters [1,21,52].
Guide Evolution Through Foresight and Adaptive Policy	Proactively steer technological trajectories through flexible governance.	Regulatory sandboxes; risk-based regulation (e.g., EU AI Act); scenario planning; multi-stakeholder engagement; sunset clauses [41,94,97,98].
Reform Education to Cultivate Competencies	Prepare individuals and workforces for effective human–machine collaboration.	Updated curricula covering human-centric design and AI ethics; upskilling and retraining programs; emphasis on inquiry-based and critical thinking skills [16,71].

Before elaborating each principle, their intended contribution should be made explicit. The five principles deliberately consolidate a consensus that spans AI ethics [20,21], human-centered computing [31], and international governance instruments [14,97]; their restatement is the starting point of this section, not its result. What the integration perspective adds is twofold. First, the content of each principle changes as integration deepens: Figure 4 maps the five principles onto the four levels of Section 2 and shows how the same principle translates into different design obligations and governance instruments at each level. Second, bodily and body-proximal integration (Levels 2 and 3) imposes demands that generic AI ethics frameworks do not cover. Where technologies cross the bodily boundary, preserving agency ceases to be only a matter of human oversight over decisions and becomes a matter of cognitive liberty and mental integrity, that is, protection of the neural substrate of decision-making itself [8,22]; defining benefit must confront enhancement rather than merely service quality; and inclusion must address coerced adoption, the pressure to augment in order to remain employable, which adds a distinct bodily-autonomy dimension to forms of coercive adoption that also arise in Level 4b software deployments. The neurorights literature [8,22,99] accordingly functions here as the template for what integration-specific ethics requires beyond existing AI frameworks.

The cells of Figure 4 are coded by provenance, so that a reader can check the contribution instead of taking it on trust. Cells marked E (established) restate an obligation already explicit in the literature cited for that level. Cells marked T (transferred) state an obligation that is established in an adjacent level or domain but is not, to our knowledge, stated as a requirement in the home literature of the level to which it is applied here. Cells marked N (new) are proposed by this review. Twelve of the twenty cells are established, five are transferred, and three are new. The three new cells are: protection and parity for those who decline augmentation, together with an explicit non-coercion duty on deployers (inclusion at Level 3); a duty to maintain usable non-algorithmic alternatives to essential services, which follows directly from the 4a–4c decomposition (inclusion at Level 4); and the capacity for informed consent to integration treated as an educational and not merely a procedural requirement (education at Level 3). The five transferred cells are: a right to remove and disconnect a worn system, which carries over to wearables the right to disconnect familiar from several European labor codes and which wearable evaluation itself does not state (agency at Level 2); demonstrated therapeutic value with an explicit boundary against enhancement (benefit at Level 3); affordable access to augmentation together with support for non-users (inclusion at Level 2); safe-use training and operator competence (education at Level 2); and civic AI literacy with participatory competence (education at Level 4). The foresight row contains no new cells: it consolidates existing governance instruments, and we claim no novelty for it. We state the distribution in this form deliberately. It bounds what the figure claims, and it lets a reader contest a particular cell against the literature cited for that level in Section 2 and for that principle here, instead of contesting the figure as a whole. The three new cells are not the whole of what Figure 4 asserts, however. Its main claim is the level-dependence itself: that the same principle changes content as integration deepens, so that an obligation discharged correctly at one level can be discharged correctly and still fail at another.

Preserve Human Agency and Rights. No matter how autonomous or intelligent machines become, humans should retain meaningful control and accountable authority in consequential domains. Santoni de Sio and van den Hoven [63] have provided a philosophical account of meaningful human control over autonomous systems, arguing that such control requires both tracing conditions (the system’s behavior must be traceable to a human’s reasons) and tracking conditions (the system must track relevant moral reasons). Wu et al. [95] survey human-in-the-loop machine learning approaches that keep human judgment integrated into AI workflows, while Endsley [6]

draws on decades of research to warn that poorly designed automation can actually undermine the very human capabilities it is meant to support, through mechanisms such as loss of situation awareness, skill degradation, and complacency. Automated systems must therefore be designed to augment human judgment without supplanting it, implementing capabilities for human override, setting hard constraints on AI behavior (e.g., forbidding actions that violate human rights or safety), and possibly instituting legal requirements for human authorization of critical AI decisions. The right to refuse or opt out of technology (such as disconnecting from an AI system or choosing a manual mode) must be upheld to protect individual freedom. Additionally, human-centric integration will benefit from recognizing that responsibility lies with human operators and organizations deploying AI, not the AI itself: legal frameworks should prevent any autonomy gap where harmful actions might escape accountability by blaming the machine [94].

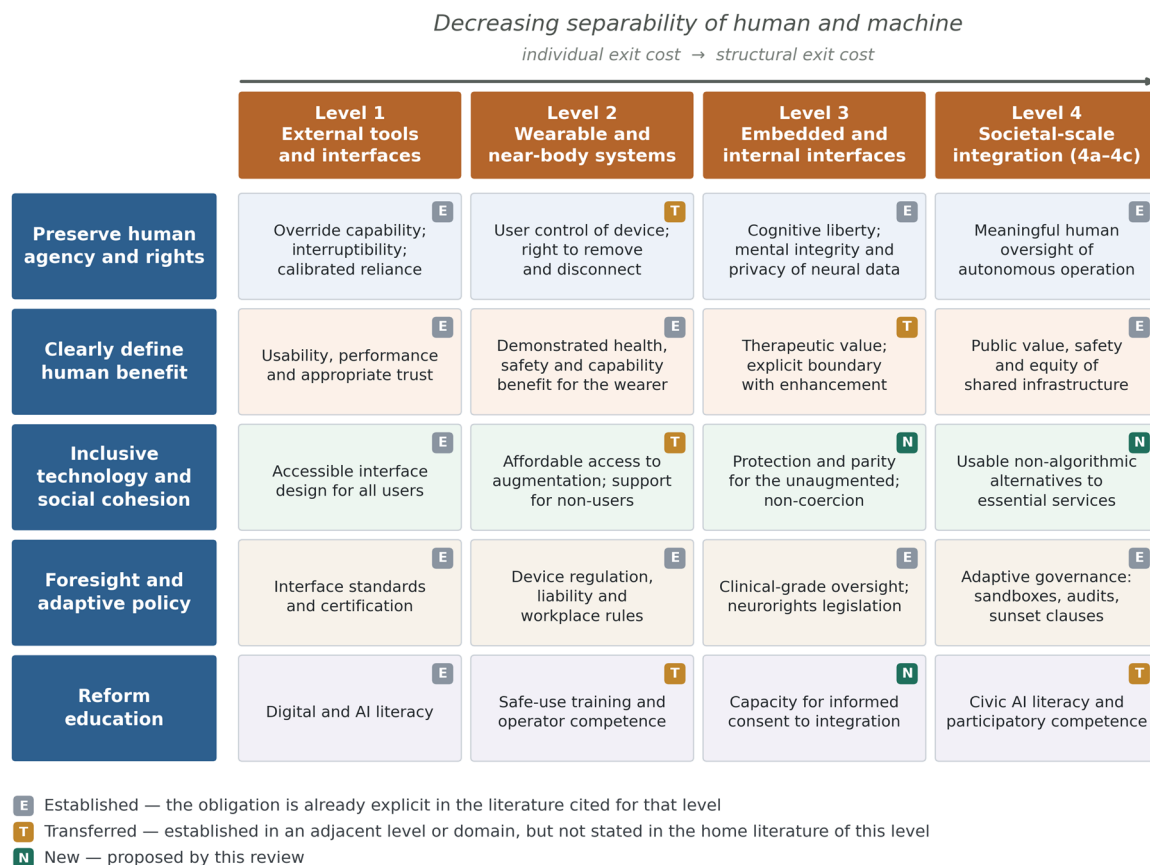


Figure 4. Mapping of the five human-centric principles onto the four levels of integration, indicating the dominant design obligations and governance instruments through which each principle is operationalized at each level. Cell entries synthesize the level-specific considerations discussed in Sections 2 and 4; representative sources: agency and rights [6,8,22,32,63,99]; human benefit [7,9,31–35,39,54]; inclusion and cohesion [1,21,52,90,93]; foresight and adaptive policy [41,54,59,94,97]; education [16,22,35,71,88]. Cells are coded by provenance: E, established in the literature cited for that level; T, transferred from an adjacent level or domain; N, proposed by this review. The coding is explained and the new cells are enumerated in the text above.

Clearly Define Human Benefit. As we design and regulate advanced HMI technologies, it is essential to ground our efforts in a clear vision of what constitutes benefit for humans. Floridi et al. [20] proposed an ethical framework for a good AI society organized around four traditional bioethical principles (beneficence, non-maleficence, autonomy, justice) plus a fifth: explicability, the requirement that AI systems be transparent and understandable. Policymakers, technologists, and society at large must engage in dialogue to define the values and outcomes we seek (for example, improved health and longevity, greater safety, enhanced capabilities, better quality of life, and respect for human dignity). Bogert et al. [100] have demonstrated experimentally that humans rely more on algorithms than social influence as tasks become more difficult, providing empirical evidence that task difficulty has measurable consequences for reliance behavior. These definitions should be informed by interdisciplinary research (including ethics, sociology, and psychology) to avoid narrow or short-term metrics. Van Wynsberghe [96] extends this argument to sustainability, emphasizing that AI should not only serve human

well-being but also ecological well-being, since the two are ultimately inseparable. By setting well-founded value-based goals for technology, we create a benchmark against which to measure the success or failure of HMI innovations, assessing systems not merely on technical performance but on their contribution to human flourishing.

Ensure Inclusive Technology and Social Cohesion. A human-centric approach insists that the benefits of human-machine integration be broadly shared across society. Not everyone will adopt or afford advanced enhancements; some may opt out of invasive integrations due to personal or cultural values. It is imperative that those who remain unaugmented are not systematically disadvantaged. Concretely, this means maintaining alternative, accessible modes of interaction with essential services that do not require advanced personal technology. For instance, if banking or public services become available through neural interfaces or augmented reality, there must still be accessible options for citizens without those tools. Policies should protect the rights and roles of individuals who may not use certain enhancements (for example, protecting unenhanced citizens as hiring of very capable cyborg-enhanced workers increases). Jobin et al. [21] reviewed the global landscape of AI ethics guidelines and found that fairness, non-discrimination, and justice are among the most frequently cited principles, yet implementation remains a challenge. The persistent pattern in which technological revolutions initially widen rather than narrow social inequalities [1,52] suggests that proactive policy intervention is necessary. Avoiding a digital divide in the era of human augmentation might require publicly funded programs to provide baseline augmentative devices (akin to public healthcare providing prosthetics), or policies that secure a guaranteed high standard of living regardless of labor market competitiveness. In the most advanced integration scenarios, non-augmented individuals might need protections analogous to disability rights legislation, ensuring they are not coerced or indirectly forced into augmentation for lack of viable non-augmented alternatives.

Guide Evolution Through Foresight and Adaptive Policy. The range of plausible futures for human-machine relations (from dystopian to utopian) points to the need for vigilant strategic foresight. Policymakers should systematically explore scenarios of technology's development (e.g., human-AI competition vs. cooperation vs. full integration) and use these to inform present-day decisions. Taeihagh [41] has emphasized that governance of AI requires adaptive approaches that can respond to rapid technological change through mechanisms such as regulatory sandboxes, sunset clauses for regulations that are regularly reviewed, and ongoing stakeholder engagement. The European Union's AI Act [97] represents a landmark attempt at comprehensive, risk-based AI regulation, classifying AI applications by risk level and imposing graduated requirements from transparency obligations to outright prohibitions. Dignum [94] has argued for a responsible AI framework that goes beyond compliance to embed ethical considerations into every stage of the AI lifecycle, from design to deployment to decommissioning. Stahl et al. [98] have empirically examined how organizations respond to the ethical issues of AI, identifying structural, procedural, and relational organizational practices needed to operationalize ethical AI principles, while Mhlanga [101] proposes a governance framework for responsible Industry 4.0 AI and identifies persistent gaps between principle and practice across jurisdictions. If tight human-AI cooperation is deemed the ideal, governments might invest in R&D for collaborative AI systems and reject AI applications that entirely replace human judgment. If integration is likely to increase, then early frameworks for certifying and monitoring augmentative technologies should be put in place now (much as biomedical devices are regulated). A key element of this principle is flexibility: regulations and institutions must be ready to evolve as technology does, avoiding both the risks of premature regulation that stifles beneficial innovation and the risks of regulatory lag that allows harmful applications to become entrenched.

Reform Education to Cultivate Human-Machine Competencies. Beyond technology design and governance, reforming educational systems is equally essential for human-centric integration. As the nature of work, governance, and social interaction is transformed by AI, curricula and training programs must be updated to equip people with the competencies needed in an AI-driven world. This includes knowledge of human-centric design and human-machine integration, the rights and governance of autonomous entities, the ethical implications of enhanced versus unenhanced humans, and an understanding of the risks posed by AI systems with extreme objective-reaching capacities. Just as important, it also means shifting the focus of education from simply answering questions and executing predefined tasks to learning how to formulate the right questions and specify desired outcomes, a skillset vital for working effectively alongside intelligent machines [16,71]. The World Economic Forum [71] and the Stanford AI Index [16] both stress that aligning workforce skills and education with rapid technological change is urgent, as AI-driven job displacement disproportionately affects workers without access to upskilling opportunities. Strategic investments in education and retraining are widely seen as necessary to ensure that no one is left behind as AI transforms jobs and society. By proactively reforming education to cultivate these new forms of knowledge and inquiry, we can better uphold human agency and ensure an inclusive, human-centric future in the face of accelerating innovation.

5. What the Framework Adds: Implications for Each Research Community

The test of a cross-level framework is whether it tells a researcher working at one level something their own literature does not. This section applies that test to each of the four communities the review addresses. The implications are deliberately narrow: each changes what should be reported, measured, or regulated, and each can be adopted or rejected on its own terms, without accepting the framework as a whole.

Wearable and industrial HMI research (Level 2): report the control regime, not only the device. Evaluations of exoskeletons and worn assistive systems characterize biomechanical load, comfort, and acceptance for the wearer in isolation [33,34]. The compound-risk analysis of Section 2 implies that the exposure a device has been validated for is not a property of the device but of the device inside a control regime: the same exoskeleton may produce a different cumulative load when the work rate is set by the wearer, by a supervisor, or by a Level 4b scheduling system optimizing throughput. Device evaluations should therefore report the pacing authority under which measurements were taken, and validation claims should be stated as conditional on it. We did not identify such reporting among the literature reviewed here, so results obtained under self-paced conditions may be generalized to algorithmically paced deployments without the generalization having been tested.

Brain–computer interface and neurotechnology research (Level 3): specify non-coercion at the level of the deployer, not the device. Neurorights are formulated principally as protections attaching to the individual and to the neural signal: cognitive liberty, mental privacy, mental integrity, psychological continuity [8,22,99]. At Level 3 these protections can be fully satisfied at the device level and still fail in practice, because the binding pressure arrives from Level 4b. A prohibition on non-consensual neural access does not constrain an employer or an insurer who conditions access to a role or a benefit on nominally voluntary adoption. Consent instruments and neurorights legislation should therefore carry a deployer-facing non-coercion clause specifying what may not be conditioned on adoption, and studies involving workplace or other non-clinical populations should report whether participants had a materially equivalent non-augmented alternative. We did not identify either practice as routine among the literature reviewed here.

AI governance and science and technology studies (Level 4): add an alternative-availability test to risk-based regulation. Risk-based instruments such as the EU AI Act [97] classify systems by application domain and by the severity of potential harm. The separability decomposition supplies a second criterion, largely orthogonal to the first, auditable, and currently absent: whether a usable non-algorithmic pathway remains available for the function the system performs. A low-risk system at 4c can bind a population more tightly than a high-risk system at 4a, because the former admits no exit. We propose alternative availability as a distinct regulatory test, comprising an obligation on the deploying authority to demonstrate that a non-algorithmic route exists and is actually used, and a trigger for heightened oversight where it does not. This also gives adaptive governance [41] a quantity to monitor over time, in place of periodic qualitative review.

Human–computer integration research (Levels 1 and 2): treat exit cost as a design variable alongside bodily agency and ownership. The human–computer integration agenda has developed a rich design space around bodily agency and bodily ownership [17–19], both properties of the coupling between a person and an artifact. Whether integration remains human-centric, however, is determined at least as much by a property that resides in no artifact at all: the availability and cost of disengagement in the setting where the artifact is deployed. A system can score well on bodily agency and still be experienced as coercive if declining it is not a live option. Reporting exit cost as a design parameter, that is, who can end the coupling, at what cost, and whether the answer changes when the system is deployed institutionally instead of voluntarily, would make integration designs comparable across the settings in which they are actually used.

The four implications share a structure. The home literature in each case specifies the coupling; the framework specifies the conditions of exit. Each community already holds the relevant value commitment; what it lacks is a reporting or regulatory variable through which that commitment can be checked. We would ask the framework to be judged on that rather than on the breadth of the synthesis.

6. Discussion: Ensuring a Human-Centric Trajectory

The principles articulated above (agency, benefit, inclusion, foresight, and education) provide a normative compass. However, translating these principles into practice requires concrete strategies in at least four interconnected domains: technology design, policy and regulation, education, and public engagement. The transition from human–machine interaction to deeper integration is not only a technological evolution, but a social one, and it demands both innovation in design and innovation in governance. The concept of human-centric integration echoes the ideals of Industry 5.0 [13] and human-centric AI (HCAI), which calls for reclaiming technology for human ends by systematically embedding human values, capabilities, and oversight into every stage

of AI system design [102]. Coeckelbergh [91] argues that AI ethics must move beyond abstract principles to engage with the concrete political, social, and relational contexts in which technologies are developed and used. Achieving this demands unprecedented cooperation between technologists, policymakers, ethicists, educators, and the public.

First, design principles for future HMI must incorporate ethics and human factors from the ground up. Umbrello [103] has demonstrated how value-sensitive design (VSD) can be applied to emerging technologies, ensuring that features like user consent, transparency, and fail-safes for manual control are embedded from the earliest stages of development. Friedman and Hendry [104] provide a comprehensive VSD methodology that integrates conceptual, empirical, and technical investigations to systematically account for human values throughout the design process. The application of VSD to HMI is particularly important because the values at stake (autonomy, privacy, dignity, safety) are not merely desirable features but fundamental human rights that must be protected by design rather than addressed retroactively. Long et al. [105] recently examined VSD applications in health AI specifically, finding that while individual-level values such as trust and autonomy receive substantial attention, organizational and societal values remain underrepresented in applied health AI design. However, Sadek and Mougenot [106] have documented persistent challenges in implementing VSD in practice more broadly: interviews with AI practitioners reveal gaps between theoretical frameworks and the organizational realities of fast-paced AI development, including insufficient stakeholder engagement, competing commercial pressures, and a lack of standardized methods for translating abstract values into concrete design requirements. Research in explainable AI (XAI) is equally important for supporting human understanding and control: Arrieta et al. [107] offer a comprehensive survey of XAI concepts and taxonomies, while Samek et al. [108] review methods for explaining deep neural networks and assess their effectiveness for building trust in human–AI systems. As HMI systems become more complex, the demand for interpretable and transparent AI will only grow, making XAI research essential for maintaining the principle of human agency articulated in Section 4. A recent meta-survey by Saeed and Omlin [109] confirms that while XAI has made significant progress in post-hoc explanation methods, persistent challenges remain in developing explanations that are both technically rigorous and meaningful to diverse end-users, particularly in high-stakes domains where human–AI collaboration is most critical. Adadi and Berrada [110] further provide a broad survey of XAI methods and applications, documenting the promise of interpretable models alongside persistent challenges in producing explanations that are useful to end-users across high-stakes settings.

This technical research must go hand in hand with social research. As Tsvetkova et al. [12] argue, a new sociology of humans and machines is needed to understand how algorithms, robots, and humans interact in networks and organizations. Rahwan et al. [11] similarly call for the study of machine behavior as an interdisciplinary endeavor that examines machines not just as engineered artifacts but as agents that participate in and shape social systems. Interdisciplinary efforts can illuminate how trust forms (or erodes) when humans work with AI side by side [32,43], or how the presence of machine agents affects human collective decision-making. Such research is essential for informing the design of HMI systems that function well not only at the individual level but also within the broader social and organizational contexts in which they are deployed.

Second, the policy environment should incentivize human-centric outcomes. The EU AI Act [97] provides a regulatory model that classifies AI systems by risk level and imposes transparency and conformity requirements, representing the most comprehensive attempt to date at binding AI regulation. UNESCO's Recommendation on the Ethics of AI [14] further establishes an international ethical benchmark centered on human dignity, agency, and community. Reddy et al. [111] propose a governance model for the application of AI in healthcare that operationalizes principled AI ethics into enforceable oversight mechanisms, while Smuha [112] offers a comparative analysis of how different regulatory traditions balance innovation incentives with ethical safeguards, framing emerging jurisdictional dynamics as a "race to AI regulation". In 2025, UNESCO extended its normative reach to neurotechnology specifically, adopting the first global Recommendation on the Ethics of Neurotechnology [99], which situates brain–computer interfaces within a human rights framework and calls upon states to prevent harmful uses including coercive control and unlawful surveillance. Ethical guidelines for brain–machine interfaces should be developed now, articulating requirements for user consent, data privacy (especially for neural data), and rigor in clinical testing. Yuste et al. [8] have called for privacy of neural data to be treated as a fundamental right, while Ienca and Andorno [22] advocate for new rights to cognitive liberty and mental integrity. On a broader scale, maintaining human control might require new legal categorizations: licensing schemes for deploying advanced AI (similar to how we license physicians, pilots, or potentially dangerous products), where organizations must demonstrate that a qualified human is overseeing high-risk AI systems at all times. Legal doctrines may also need updates to handle integrated human–machine entities (for example, confirming that augmented humans bear responsibility for the actions of their enhancements, and that AI components have no independent rights or

liabilities). On the flip side, those who remain less integrated might need protections analogous to disability rights, to ensure they are not coerced or indirectly forced into augmentation for lack of non-augmented options.

Lastly, public engagement and inclusivity must drive the conversation about HMI's future. The form that human-machine integration ultimately takes will reflect societal choices about what we value. It is essential to democratize this process by involving citizens in dialogues about new technologies and their impacts. Public consultations, participatory technology assessments, and transparency about the capabilities and risks of AI can help build a shared understanding. Gillath et al. [88] have shown empirically that attachment styles shape patterns of trust in AI, suggesting that how AI capabilities are communicated and governed can substantially influence public acceptance of technology integration. This inclusive approach can also allay public fears by showing that strong governance structures are in place to keep technology a tool for human benefit, not an uncontrollable force. By embedding humanistic values into both technology design and regulation, we can approach the coming integration with our eyes open.

7. Conclusions

The evolution from basic human-machine interaction to ever deeper human-machine integration represents a transformative frontier for society. We have unprecedented opportunities to enhance human abilities, from AI-enabled exoskeletons that may reduce physical strain and injury risk [33,34] to brain interfaces that restore lost functions [7,38] to intelligent infrastructure that can improve the quality of urban life [39,40]. At the same time, the risks of misalignment or misuse are stark: without guiding principles, we could face a future where technology erodes personal autonomy, blurs accountability, or deepens social inequalities [42,70]. Ensuring a human-centric future for HMI is thus a multidimensional challenge, involving technology design, ethical norms, economic policy, education, and law all at once. The overlapping strategies outlined in this review reflect that complexity. They urge us to place human reasoning, values, and well-being at the core of every step toward integration. In practical terms, this means always asking: Does a given human-machine system advance human welfare and uphold rights? Is it preserving human agency and dignity? If the answer is uncertain, that system requires redesign or regulation before widespread adoption.

In summary, a human-centric approach to human-machine integration demands: clear definitions of the human purposes technology should serve [20,102]; rigorous protection of human agency and oversight in all AI and interface deployments [6,60,61,63,95]; inclusive strategies so that technological advances benefit all of humanity, not only tech adopters or elites [21]; continuous foresight to anticipate future shifts and steer them responsibly [41,94]; and reformed education systems that cultivate the competencies needed to collaborate effectively with intelligent machines and govern their deployment wisely [16,71]. The concept of human-centric is not a vague slogan but a concrete design and policy mandate: technology must be a means to human ends. The coming decades will test our collective ability to uphold this mandate, especially as the lures of efficiency, profit, or even transhuman enhancement arise [90]. By adhering to fundamental humanistic principles and proactively crafting policies and technologies that reinforce them, we can ensure that even as we integrate more intimately with our machines, we remain the authors of our destiny and the primary beneficiaries of progress. Concretely, the framework asks four communities for four additions to current practice: the pacing authority under which worn systems are validated, a deployer-facing non-coercion clause in neurorights instruments, an alternative-availability test alongside risk-based AI regulation, and exit cost reported as a design variable in integration research (Section 5). Realizing this vision, however, will require sustained research effort to address the gaps and open questions that remain in the field, which we outline below.

8. Prospects

Despite the growing body of research surveyed in this review, several notable gaps remain that define the frontier of future inquiry. The vast majority of empirical studies on human-machine teaming have been conducted in controlled laboratory settings with short-duration tasks [72]; longitudinal field studies that track how trust, reliance, and skill dynamics evolve over months or years of sustained human-machine integration are largely absent. The ethical and governance frameworks reviewed here [14,20,21,97] remain predominantly Western and developed-world-centric; it is unclear how well they translate to different cultural, economic, and institutional contexts, particularly in the Global South where the benefits and risks of integration technologies may be distributed very differently. The literature on neurotechnology ethics [8,22,99] has advanced primarily at the level of principle rather than implementation; practical frameworks for auditing neural data use, certifying BCI safety, or adjudicating disputes arising from augmented cognition remain underdeveloped, a gap that the rapid commercialization of BCIs [54] makes increasingly urgent. While the Industry 5.0 discourse [13,79,80,82,83] has

generated valuable normative frameworks, empirical evidence on whether human-centric manufacturing designs actually improve worker well-being (as opposed to merely productivity) is still sparse. Finally, the interactions between different levels of integration (for example, how societal-scale AI infrastructure (Level 4) shapes the choices and constraints available to individuals at the wearable (Level 2) or embedded (Level 3) levels) have received little systematic attention, despite their importance for understanding the full dynamics of an increasingly integrated technological ecosystem.

Building on these observations, several concrete research directions deserve particular attention (see Figure 5). First, longitudinal empirical studies are needed to examine how prolonged use of augmentative technologies (from wearable exoskeletons to BCIs) affects human cognition, well-being, and social relationships. The human factors literature has documented how reliance on automation can lead to skill degradation and loss of situation awareness over time [6,42]; similar dynamics may emerge with more intimate forms of human–machine integration, and understanding these effects will require sustained observation across diverse populations and contexts. The emerging concept of Human Digital Twins [85,86] offers a promising methodological tool for such research, enabling continuous monitoring and prediction of individual workers’ physiological and cognitive states during human–machine interaction. The research agenda outlined by Wang et al. [51] for human-centric assembly in smart factories, spanning augmented robots, cognitive systems, mixed reality, and brain–robotic controls, provides a concrete roadmap for empirical investigation in manufacturing contexts.

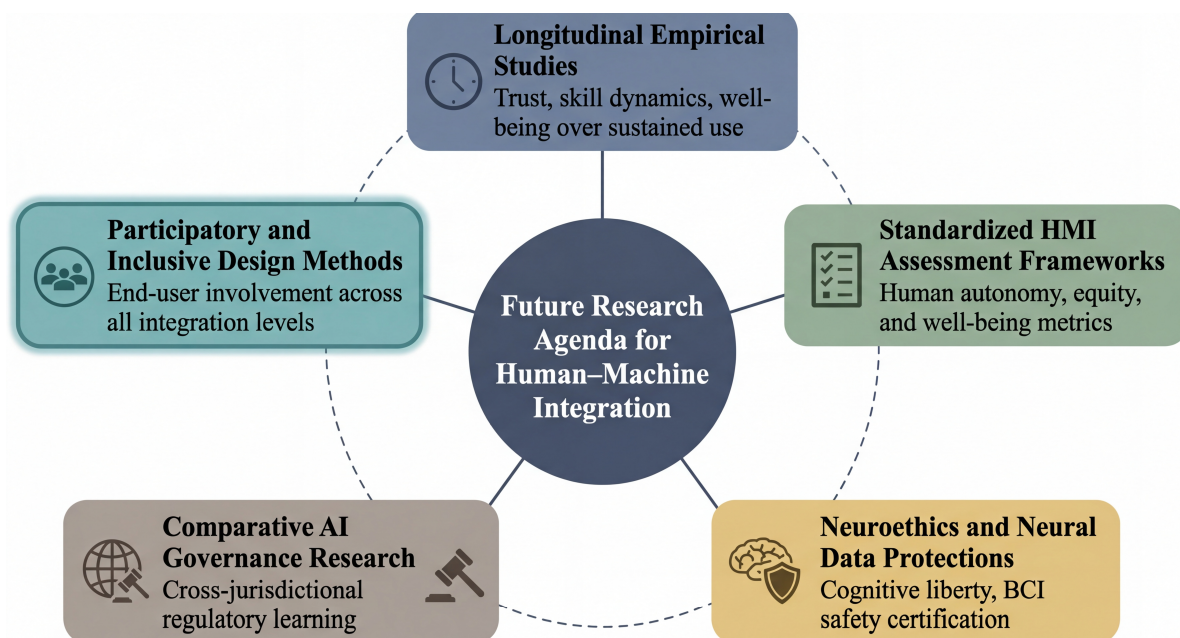


Figure 5. Future research agenda for human–machine integration, highlighting five interconnected research directions. Participatory and inclusive design methods (left) serve as a cross-cutting theme that informs all other directions.

Second, interdisciplinary collaborations between engineers, ethicists, cognitive scientists, and policymakers should develop standardized assessment frameworks for evaluating whether HMI systems truly serve human-centric goals. Drawing on empirical findings about reliance on algorithmic advice under task difficulty [100] and value-sensitive design methodology [104], such frameworks could provide practical tools for assessing the impact of HMI technologies on human autonomy, well-being, social connectedness, and equity, going beyond narrow technical performance metrics to capture the full spectrum of human values at stake.

Third, as the EU AI Act [97] enters broad application and enforcement, other jurisdictions are developing their own regulatory approaches. Annex III high-risk rules apply from 2 December 2027, whereas rules for high-risk systems embedded in regulated products apply from 2 August 2028. Comparative governance research will be essential to identify best practices for adaptive AI regulation. The tension between regulatory harmonization and the need for context-sensitive approaches that account for different cultural values, economic conditions, and institutional capacities will be a central challenge [41,94,112]. Experimental governance approaches (such as regulatory sandboxes, algorithmic auditing requirements, and multi-stakeholder governance bodies) deserve rigorous evaluation. Guio Español and Koenig [113] have proposed a learning-centric model for AI regulatory sandboxes that foregrounds iterative policy learning, arguing that sandbox design must be adapted to different institutional and economic contexts rather than simply transplanted from high-income jurisdictions.

Fourth, the growing capabilities of neurotechnologies call for urgent development of neuroethics frameworks and legal protections for neural data privacy [8,22]. As BCIs move from laboratory to clinical and potentially consumer applications, the questions of who owns neural data, how it may be used, and what safeguards protect cognitive liberty will become pressing practical concerns rather than speculative thought experiments. UNESCO's 2025 Recommendation on the Ethics of Neurotechnology [99] represents a significant first step, but translating its principles into binding national legislation remains an open challenge [54,55].

Finally, participatory design methods that involve diverse end-users, including those who may choose not to adopt new technologies, should become standard practice in HMI research and development. The insights from the sociology of human–machine relations [11,12] and the growing body of research on human–AI teaming dynamics [73–75] can inform more inclusive design processes that account for the full range of ways people may interact with (or choose to remain separate from) integrated technological systems.

Author Contributions

S.G.: conceptualization, methodology, writing—original draft preparation, writing—reviewing and editing; T.W.: conceptualization, methodology, writing—original draft preparation, writing—reviewing and editing; X.V.W.: supervision, writing—reviewing and editing; L.W.: supervision, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was supported by the Centre of Excellence in Production Research (XPRES), a joint initiative of RISE AB, KTH Royal Institute of Technology, and Mälardalen University.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments

This research was performed in the context of the XPRES environment at RISE AB, KTH, and Mälardalen University.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used ChatGPT (GPT-5.5) and Claude (Opus 4.8) to polish the manuscript's language and grammar. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Korinek, A.; Stiglitz, J.E. Artificial Intelligence, Globalization, and Strategies for Economic Development. *Natl. Bur. Econ. Res. Work. Pap.* **2021**, 28453. <https://doi.org/10.3386/w28453>.
2. Lane, M.; Williams, M.; Broecke, S. The Impact of AI on the Workplace: Main Findings from the OECD AI Surveys of Employers and Workers. *OECD Soc. Employ. Migr. Work. Pap.* **2023**, 288, 156.
3. Kurzweil, R. *The Singularity is Near: When Humans Transcend Biology*; Viking: New York, NY, USA, 2005.
4. Kurzweil, R. *The Singularity Is Nearer: When We Merge with AI*; Viking (Penguin Books): New York, NY, USA, 2024.
5. Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. A Model for Types and Levels of Human Interaction with Automation. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2000**, 30, 286–297.

6. Endsley, M.R. From Here to Autonomy: Lessons Learned from Human–Automation Research. *Hum. Factors* **2017**, *59*, 5–27.
7. Lebedev, M.A.; Nicolelis, M.A.L. Brain–Machine Interfaces: From Basic Science to Neuroprostheses and Neurorehabilitation. *Physiol. Rev.* **2017**, *97*, 767–837.
8. Yuste, R.; Goering, S.; Agüera y Arcas, B.; et al. Four Ethical Priorities for Neurotechnologies and AI. *Nature* **2017**, *551*, 159–163.
9. Saha, S.; Mamun, K.A.; Ahmed, K.; et al. Progress in Brain Computer Interface: Challenges and Opportunities. *Front. Syst. Neurosci.* **2021**, *15*, 578875.
10. Vaccaro, M.; Almaatouq, A.; Malone, T. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nat. Hum. Behav.* **2024**, *8*, 2293–2303.
11. Rahwan, I.; Cebrian, M.; Obradovich, N.; et al. Machine Behaviour. *Nature* **2019**, *568*, 477–486.
12. Tsvetkova, M.; Yasseri, T.; Pescetelli, N.; et al. A New Sociology of Humans and Machines. *Nat. Hum. Behav.* **2024**, *8*, 1864–1876.
13. Breque, M.; De Nul, L.; Petridis, A. *Industry 5.0: Towards a Sustainable, Human-Centric and Resilient European Industry*; European Commission: Brussels, Belgium, 2021.
14. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021.
15. IMF. *New Skills and AI Are Reshaping the Future of Work*; International Monetary Fund: Washington, DC, USA, 2026.
16. Maslej, N.; Fattorini, L.; Perrault, R.; et al. *The AI Index 2025 Annual Report*; Stanford Institute for Human-Centered AI: Stanford, CA, USA, 2025.
17. Farooq, U.; Grudin, J. Human–Computer Integration. *Interactions* **2016**, *23*, 26–32.
18. Mueller, F.F.; Lopes, P.; Strohmeier, P.; et al. Next Steps for Human–Computer Integration. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20), Honolulu, HI, USA, 25–30 April 2020; pp. 1–15.
19. Mueller, F.F.; Semertzidis, N.; Andres, J.; et al. Human–Computer Integration: Towards Integrating the Human Body with the Computational Machine. *Found. Trends Hum.–Comput. Interact.* **2022**, *16*, 1–64.
20. Floridi, L.; Cowls, J.; Beltrametti, M.; et al. AI4People: An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707.
21. Jobin, A.; Ienca, M.; Vayena, E. The Global Landscape of AI Ethics Guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399.
22. Ienca, M.; Andorno, R. Towards New Human Rights in the Age of Neuroscience and Neurotechnology. *Life Sci. Soc. Policy* **2017**, *13*, 5.
23. Grant, M.J.; Booth, A. A Typology of Reviews: An Analysis of 14 Review Types and Associated Methodologies. *Health Inf. Libr. J.* **2009**, *26*, 91–108.
24. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71.
25. Greenhalgh, T.; Thorne, S.; Malterud, K. Time to Challenge the Spurious Hierarchy of Systematic over Narrative Reviews? *Eur. J. Clin. Invest.* **2018**, *48*, e12931.
26. Onnasch, L.; Roesler, E. A Taxonomy to Structure and Analyze Human–Robot Interaction. *Int. J. Soc. Robot.* **2021**, *13*, 833–849.
27. Dellermann, D.; Ebel, P.; Söllner, M.; et al. Hybrid Intelligence. *Bus. Inf. Syst. Eng.* **2019**, *61*, 637–643.
28. Licklider, J.C.R. Man–Computer Symbiosis. *IRE Trans. Hum. Factors Electron.* **1960**, *1*, 4–11.
29. Jarrahi, M.H. Artificial Intelligence and the Future of Work: Human–AI Symbiosis in Organizational Decision Making. *Bus. Horiz.* **2018**, *61*, 577–586.
30. Liu, S.; Wang, L.; Wang, X.V. Symbiotic Human–Robot Collaboration: Multimodal Control Using Function Blocks. *Procedia CIRP* **2020**, *93*, 1188–1193.
31. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
32. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. *Hum. Factors* **2004**, *46*, 50–80.
33. de Looze, M.P.; Bosch, T.; Krause, F.; et al. Exoskeletons for Industrial Application and Their Potential Effects on Physical Work Load. *Ergonomics* **2016**, *59*, 671–681.
34. Crea, S.; Beckerle, P.; De Looze, M.; et al. Occupational Exoskeletons: A Roadmap Toward Large-Scale Adoption. *Wearable Technol.* **2021**, *2*, e11.
35. Romero, D.; Stahre, J.; Wuest, T.; et al. Towards an Operator 4.0 Typology: A Human-Centric Perspective on the Fourth Industrial Revolution Technologies. In Proceedings of the 46th International Conference on Computers & Industrial Engineering (CIE46), Tianjin, China, 29–31 October 2016; pp. 1–11.
36. Musk, E.; Neuralink. An Integrated Brain–Machine Interface Platform with Thousands of Channels. *J. Med. Internet Res.* **2019**, *21*, e16194.
37. ClinicalTrials.gov. *PRIME: An Early Feasibility Study of a Precise Robotically Implanted Brain–Computer Interface for the Control of External Devices*; NCT06429735. Available online: <https://clinicaltrials.gov/study/NCT06429735> (accessed on 19 August 2026).

38. Neuralink. *A Year of Telepathy*. Available online: <https://neuralink.com/updates/a-year-of-telepathy/> (accessed on 19 August 2026).
39. Allam, Z.; Dhunny, Z.A. On Big Data, Artificial Intelligence and Smart Cities. *Cities* **2019**, *89*, 80–91.
40. Yigitcanlar, T.; Corchado, J.M.; Mehmood, R.; et al. Responsible Urban Innovation with Local Government Artificial Intelligence (AI): A Conceptual Framework and Research Agenda. *J. Open Innov. Technol. Mark. Complex.* **2021**, *7*, 71.
41. Taeihagh, A. Governance of Artificial Intelligence. *Policy Soc.* **2021**, *40*, 137–157.
42. Parasuraman, R.; Manzey, D.H. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Hum. Factors* **2010**, *52*, 381–410.
43. Hancock, P.A.; Billings, D.R.; Schaefer, K.E.; et al. A Meta-Analysis of Factors Affecting Trust in Human–Robot Interaction. *Hum. Factors* **2011**, *53*, 517–527.
44. Mehrotra, S.; Degachi, C.; Vereschak, O.; et al. A Systematic Review on Fostering Appropriate Trust in Human–AI Interaction. *ACM J. Responsible Comput.* **2024**, *1*, 1–45.
45. Dittmar, E.C.; Sposato, M. Optimising Human–AI Decision Performance: A Trust and Capability Framework for Knowledge Management. *Knowl. Process Manag.* **2026**, *33*, 757–771. <https://doi.org/10.1002/kpm.70059>.
46. Bansal, G.; Wu, T.; Zhou, J.; et al. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21) Yokohama, Japan, 8–13 May 2021; Article 81, pp. 1–16. <https://doi.org/10.1145/3411764.3445717>.
47. Glikson, E.; Woolley, A.W. Human Trust in Artificial Intelligence: Review of Empirical Research. *Acad. Manag. Ann.* **2020**, *14*, 627–660.
48. Haghighi, A.; Cheraghi, M.; Pocachard, J.; et al. A Comprehensive Review and Bibliometric Analysis on Collaborative Robotics for Industry: Safety Emerging as a Core Focus. *Front. Robot. AI* **2025**, *12*, 1605682.
49. Shah, R.; Doss, A.S.A.; Lakshmaiya, N. Advancements in AI-Enhanced Collaborative Robotics: Towards Safer, Smarter, and Human-Centric Industrial Automation. *Results Eng.* **2025**, *27*, 105704.
50. Patalas-Maliszewska, J.; Łosyk, H.; Dudek, A. Improving Safety in Human–Robot Collaboration toward Sustainable Production in Industry 5.0. *J. Intell. Manuf.* **2025**, *37*, 3069–3104.
51. Wang, L.; Gao, R.X.; Krüger, J.; et al. Human-Centric Assembly in Smart Factories. *CIRP Ann.* **2025**, *74*, 789–815.
52. Brynjolfsson, E.; McAfee, A. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*; W.W. Norton: New York, NY, USA, 2014.
53. Shin, Y. Toward Human-Centered Artificial Intelligence for Users’ Digital Well-Being: Systematic Review, Synthesis, and Future Directions. *JMIR Hum. Factors* **2025**, *12*, e69533.
54. Boonstra, J.T. Ethical Imperatives in the Commercialization of Brain–Computer Interfaces. *IBRO Neurosci. Rep.* **2025**, *19*, 718–724.
55. Ienca, M.; Haselager, P. Hacking the Brain: Brain–Computer Interfacing Technology and the Ethics of Neurosecurity. *Ethics Inf. Technol.* **2016**, *18*, 117–129.
56. OECD. *Smart Cities and Inclusive Growth*; OECD Publishing: Paris, France, 2020.
57. Zheng, P.; Xia, L.; Li, C.; et al. Towards Self-X Cognitive Manufacturing Network: An Industrial Knowledge Graph-Based Multi-Agent Reinforcement Learning Approach. *J. Manuf. Syst.* **2021**, *61*, 16–26.
58. Yigitcanlar, T.; Kankanamge, N.; Regona, M.; et al. Artificial Intelligence Technologies and Related Urban Planning and Development Concepts: How Are They Perceived and Utilized in Australia? *J. Open Innov. Technol. Mark. Complex.* **2020**, *6*, 187.
59. Aloisi, A.; De Stefano, V. *Your Boss Is an Algorithm: Artificial Intelligence, Platform Work and Labour*; Hart Publishing: Oxford, UK, 2022.
60. European Data Protection Supervisor. *Human Oversight of Automated Decision-Making*; TechDispatch No. 2/2025; EDPS: Brussels, Belgium, 2025.
61. Wood, N.G. Autonomous and AI-Enabled Systems: Extensions or Replacements of Human Will and Control? *Ethics Inf. Technol.* **2026**, *28*, 3.
62. Brynjolfsson, E.; Li, D.; Raymond, L.R. Generative AI at Work. *Q. J. Econ.* **2025**, *140*, 889–942.
63. Santoni de Sio, F.; van den Hoven, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Front. Robot. AI* **2018**, *5*, 15.
64. Raisch, S.; Krakowski, S. Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Acad. Manag. Rev.* **2021**, *46*, 192–210.
65. Frey, C.B.; Osborne, M.A. The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technol. Forecast. Soc. Chang.* **2017**, *114*, 254–280.
66. Acemoglu, D.; Restrepo, P. Automation and New Tasks: How Technology Displaces and Reinstates Labor. *J. Econ. Perspect.* **2019**, *33*, 3–30.

67. Autor, D.H. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *J. Econ. Perspect.* **2015**, *29*, 3–30.
68. Autor, D.; Chin, C.; Salomons, A.; et al. New Frontiers: The Origins and Content of New Work, 1940–2018. *Q. J. Econ.* **2024**, *139*, 1399–1465.
69. Acemoglu, D. The Simple Macroeconomics of AI. *Econ. Policy* **2025**, *40*, 13–58.
70. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014.
71. World Economic Forum. *The Future of Jobs Report 2025*; WEF: Geneva, Switzerland, 2025.
72. O'Neill, T.; McNeese, N.; Barron, A.; et al. Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Hum. Factors* **2022**, *64*, 904–938.
73. Hemmer, P.; Schemmer, M.; Köhl, N.; et al. Complementarity in Human–AI Collaboration: Concept, Sources, and Evidence. *Eur. J. Inf. Syst.* **2025**, *34*, 979–1002.
74. Seeber, I.; Bittner, E.; Briggs, R.O.; et al. Machines as Teammates: A Research Agenda on AI in Team Collaboration. *Inf. Manag.* **2020**, *57*, 103174.
75. Lou, B.; Lu, T.; Raghu, T.S.; et al. Unraveling Human–AI Teaming: A Review and Outlook. *arXiv* **2025**, arXiv:2504.05755.
76. Susarla, A.; Gopal, R.; Thatcher, J.B.; et al. The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Inf. Syst. Res.* **2023**, *34*, 399–408.
77. Bender, E.M.; Gebru, T.; McMillan-Major, A.; et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) Virtual, 3–10 March 2021; pp. 610–623. <https://doi.org/10.1145/3442188.3445922>.
78. Bai, H.; Voelkel, J.G.; Muldowney, S.; et al. LLM-Generated Messages Can Persuade Humans on Policy Issues. *Nat. Commun.* **2025**, *16*, 6037. <https://doi.org/10.1038/s41467-025-61345-5>.
79. Xu, X.; Lu, Y.; Vogel-Heuser, B.; et al. Industry 4.0 and Industry 5.0: Inception, Conception and Perception. *J. Manuf. Syst.* **2021**, *61*, 530–535. <https://doi.org/10.1016/j.jmsy.2021.10.006>.
80. Leng, J.; Sha, W.; Wang, B.; et al. Industry 5.0: Prospect and Retrospect. *J. Manuf. Syst.* **2022**, *65*, 279–295.
81. Wang, L. A Futuristic Perspective on Human-Centric Assembly. *J. Manuf. Syst.* **2022**, *62*, 199–201.
82. Nahavandi, S. Industry 5.0—A Human-Centric Solution. *Sustainability* **2019**, *11*, 4371.
83. Xu, J.; Sun, Q.; Han, Q.-L.; et al. When Embodied AI Meets Industry 5.0: Human-Centered Smart Manufacturing. *IEEE/CAA J. Autom. Sin.* **2025**, *12*, 485–501.
84. Longo, F.; Padovano, A.; Umbrello, S. Value-Oriented and Ethical Technology Engineering in Industry 5.0: A Human-Centric Perspective for the Design of the Factory of the Future. *Appl. Sci.* **2020**, *10*, 4182. <https://doi.org/10.3390/app10124182>.
85. Gaffinet, B.; Al Haj Ali, J.; Naudet, Y.; et al. Human Digital Twins: A Systematic Literature Review and Concept Disambiguation for Industry 5.0. *Comput. Ind.* **2025**, *166*, 104230. <https://doi.org/10.1016/j.compind.2024.104230>.
86. Miller, M.E.; Spatz, E. A Unified View of a Human Digital Twin. *Hum.-Intell. Syst. Integr.* **2022**, *4*, 23–33.
87. Maddikunta, P.K.R.; Pham, Q.-V.; Prabadevi, B.; et al. Industry 5.0: A Survey on Enabling Technologies and Potential Applications. *J. Ind. Inf. Integr.* **2022**, *26*, 100257. <https://doi.org/10.1016/j.jii.2021.100257>.
88. Gillath, O.; Ai, T.; Branicky, M.S.; et al. Attachment and Trust in Artificial Intelligence. *Comput. Hum. Behav.* **2021**, *115*, 106607. <https://doi.org/10.1016/j.chb.2020.106607>.
89. Butler, J.; Jaffe, S.; Janssen, R.; et al. *New Future of Work Report 2025*; Technical Report MSR-TR-2025-58; Microsoft: Redmond, WA, USA, 2025.
90. Sorgner, S.L. *We Have Always Been Cyborgs: Digital Data, Gene Technologies, and an Ethics of Transhumanism*; Bristol University Press: Bristol, UK, 2022.
91. Coeckelbergh, M. *AI Ethics*; MIT Press: Cambridge, MA, USA, 2020.
92. Matthews, M.J.; Su, R.; Yonish, L.; et al. A Review of Artificial Intelligence, Algorithms, and Robots Through the Lens of Stakeholder Theory. *J. Manag.* **2025**, *51*, 2627–2676. <https://doi.org/10.1177/01492063241311855>.
93. Alegre, S.; Vosloo, S. *Neurotechnology and Children's Rights: Preparing for the Future*; UNICEF Innocenti: Florence, Italy, 2025.
94. Dignum, V. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*; Springer: Cham, Switzerland, 2019.
95. Wu, X.; Xiao, L.; Sun, Y.; et al. A Survey of Human-in-the-Loop for Machine Learning. *Future Gener. Comput. Syst.* **2022**, *135*, 364–381. <https://doi.org/10.1016/j.future.2022.05.014>.
96. van Wynsberghe, A. Sustainable AI: AI for Sustainability and the Sustainability of AI. *AI Ethics* **2021**, *1*, 213–218.
97. European Parliament; Council of the European Union. *Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. Available online: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj> (accessed on 19 August 2026).

98. Stahl, B.C.; Antoniou, J.; Ryan, M.; et al. Organisational Responses to the Ethical Issues of Artificial Intelligence. *AI Soc.* **2022**, *37*, 23–37. <https://doi.org/10.1007/s00146-021-01148-6>.
99. UNESCO. *Recommendation on the Ethics of Neurotechnology*; UNESCO: Paris, France, 2025.
100. Bogert, E.; Schecter, A.; Watson, R.T. Humans Rely More on Algorithms than Social Influence as a Task Becomes More Difficult. *Sci. Rep.* **2021**, *11*, 8028. <https://doi.org/10.1038/s41598-021-87480-9>.
101. Mhlanga, D. *Responsible Industry 4.0: A Framework for Human-Centered Artificial Intelligence*; Routledge: London, UK, 2023.
102. Schmager, S.; Pappas, I.O.; Vassilakopoulou, P. Understanding Human-Centred AI: A Review of Its Defining Elements and a Research Agenda. *Behav. Inf. Technol.* **2025**, *44*, 3771–3810. <https://doi.org/10.1080/0144929X.2024.2448719>.
103. Umbrello, S. Imaginative Value Sensitive Design: Using Moral Imagination Theory to Inform Responsible Technology Design. *Sci. Eng. Ethics* **2020**, *26*, 575–595.
104. Friedman, B.; Hendry, D.G. *Value Sensitive Design: Shaping Technology with Moral Imagination*; MIT Press: Cambridge, MA, USA, 2019.
105. Long, Y.; Novak, L.; Walsh, C.G. Searching for Value Sensitive Design in Applied Health AI: A Narrative Review. *Yearb. Med. Inform.* **2024**, *33*, 75–82. <https://doi.org/10.1055/s-0044-1800723>.
106. Sadek, M.; Mougenot, C. Challenges in Value-Sensitive AI Design: Insights from AI Practitioner Interviews. *Int. J. Hum.-Comput. Interact.* **2025**, *41*, 10877–10894. <https://doi.org/10.1080/10447318.2024.2439021>.
107. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>.
108. Samek, W.; Montavon, G.; Lapuschkin, S.; et al. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* **2021**, *109*, 247–278. <https://doi.org/10.1109/JPROC.2021.3060483>.
109. Saeed, W.; Omlin, C. Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities. *Knowl.-Based Syst.* **2023**, *263*, 110273.
110. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.
111. Reddy, S.; Allan, S.; Coghlan, S.; et al. A Governance Model for the Application of AI in Health Care. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 491–497. <https://doi.org/10.1093/jamia/ocz192>.
112. Smuha, N.A. From a ‘Race to AI’ to a ‘Race to AI Regulation’: Regulatory Competition for Artificial Intelligence. *Law Innov. Technol.* **2021**, *13*, 57–84.
113. Guio Español, A.; Koenig, P.D. Regulatory Sandboxes for AI in the Majority World: A Learning-Centric Approach to Legal Adaptation. *Camb. Forum AI Law Gov.* **2025**, *1*, e42. <https://doi.org/10.1017/cfl.2025.10015>.