HUMAN-AI COLLABORATION: BALANCING AGENTIC AI AND AUTONOMY IN HYBRID SYSTEMS

Gaurav Samdani [1], Ganesh viswanathan [2] and Abirami Dasu Jegadeesh [3]

[1] Department of Data Science, University of North Carolina, Charlotte
[2] Information Technology, 3530 Alister Ave concord 28027
[3] Department of Information Technology, Anna university, Chennai

ABSTRACT

In this article, the author explores the tension between the human factor and artificial intelligence as a symbiosis of two effective approaches to solving multifaceted, realistic tasks. Considering the premises of human-AI cooperation, it identifies how combined structures can improve these processes as decision making, scalability and flexibility in spheres including healthcare, auto transport industry as well as education.

The discussion combines theories and case studies to explain how hybrid systems may retain transparent, fair, and ethical procedures while achieving operational performance. Beneficial samples include one focusing on developing possible issues with the implementation of, for instance, human supervision of AI and the growth of AI decision making self-governance, the problem of AI biases, and others pertaining to drawbacks of over-essentialization of AI.

KEYWORDS

Human-AI collaboration, Hybrid systems, Agentic AI and autonomy,Ethical considerations,Decision-making frameworks

1. INTRODUCTION

Human-AI collaboration is a topic of increasing interest as society becomes more aware of its potential. This phenomenon is driven by increased interest in the development of AI from both industry and academia, further enabled by an ongoing increase in computing power and storage capacity. Due to the ongoing scaling of AI technologies, they are also becoming more relevant and critical in many aspects of modern society and human life: AI systems can outperform humans in solving pattern recognition tasks, help address large-scale optimization problems, and even control complex systems. One specific aspect of AI technology development is of pivotal importance in this context. In recent years, there has been a push for the abandonment of large-scale “beauty pageants,” towards instead using hybrid human-AI teams for the performance of complex cognitive tasks, as these hybrids are often reported as more effective and efficient. At the heart of these hybrid systems is a seamless and continuously reconfiguring interaction between agentic AI and AI autonomy.

One is thus motivated to study possible ways of designing these hybrid systems such that they can balance the symmetry of both components of the system to arrive at effective decision-making. One essential requirement is that the adopted organization must reflect and integrate the strengths of the individual components in such a way that the emergent behavior is consistent with collective goals. There are tangible benefits of collaboration between AI and humans, as AI

systems can provide additional expertise to humans, offer more comprehensive and faster responses to complex events, enable the evaluation of a larger set of decision options in constrained time limits, and process and analyze more information than a human operator can. However, to maximize the benefits, the different but complementary capabilities of AI and humans need to blend effectively. This is a significant challenge for the collaboration of human operators with current and future AI tools, as it requires a balance between centralized decision-making in complex multi-agent systems and decentralized control, each able to act independently as needed in their specific environment.

1.1. Background and Significance

Background and Significance Recent technological advances have resulted in a variety of collaborative systems that draw upon a mix of human and AI capabilities to equilibrate to the strengths of each constituent part. These systems can be found across a variety of domains, including healthcare, education, and the development of new technologies. For example, doctors in hospital settings often use AI to aid in clinical assessment; similarly, amateur and professional athletes wear sensors capturing a variety of muscle and body measurements, which they review to enhance their form. Technological changes have put the specialists in a varied set of disciplines who are building these systems in conversation not only with humanists and ethicists interested in the psychological, economic, and philosophical import of these newly shaped physio-computational systems but also regulators and ethicists contributing to discussions of AI policy and regulation. Medical school deans may wonder whether and how using AI to create feedback loops over their students' learning would shape education in their institutions and want to engage educational theorists in making those decisions. Technologists working to build these educational AI systems also require a varied set of tools spanning textbook knowledge and contemporary research in areas such as developmental psychology, human-computer interaction, and machine learning research about human feedback techniques. Taken together, the experiences in this complex, highly interdisciplinary space demonstrate the need for a detailed understanding of how engineering decisions lead to shifts in the distributional position of agentic AI across human and AI contributors. In addition, if these hybrid systems persist, the long-term scalability of the system through long-lived power-sharing arrangements between humans and AI remains to be seen.

1.2. Research Objectives

This research attempts to investigate hybrid human-AI collaborations that capture a number of characteristics of both autonomous and agentic AI focused systems. The goals of this work are twofold: (1) Understanding: What defines an effective and successful hybrid human-artificial intelligence interaction, especially considering interactions where AI and/or automation may be reshaping users' decisions (including actions and/or thought processes) as they occur? (2) Optimization: For those considering building hybrid human-AI systems, what is the best practice to help ensure they are optimized in terms of their interactions and impact? In addition to providing theoretical contributions to clarify a more in-depth account of effective human-AI interactions, it is suggested that the empirically derived conceptual framework will enable technological and software engineers, along with stakeholders, to cultivate and refine new tools for optimizing hybrid human-AI interactions in the future.

Research Objective 1: Understanding. The primary research question (RQ1) that is guiding this study is as follows: What defines an effective and successful hybrid human-artificial intelligence interaction, especially those interactions where AI and/or automation may be reshaping users' decisions (including actions and/or thought processes as they occur)? RQ1 seeks to disentangle the factors that contribute to the successful interplay between human and machine, and especially

the previously unexplored domain where AI may or may not be performing tasks chosen by the human designer, but impacting the user or point of sale at the time of decision making. A hybrid system in AI is likely to be more effective if this interaction is responsive to users, data, or environmental signals. Understanding the dimensions and indicators of such efficacy can better inform the design of future systems. To create a more informed and effective design practice, we must first capture the existing behaviors and mechanisms that result in positive experiences and synchronized actions between humans and AI. To consolidate the insights observed, we suggest studying a real-world commercial setting to investigate which hybrid human-artificial intelligence systems are most successful. For this, RQ2 was designed: How do hybrid systems interact with people and how does it impact behavior?

2. FOUNDATIONS OF HUMAN-AI COLLABORATION

Human-AI collaboration is based on the interaction and collaboration between human users and AI systems. This type of hybrid interaction can be more conversational or involve more explicit collaboration in decision-making, interaction with real-world components, and so on, and encompasses a wide range of terms, including mixed-initiative systems, collaborative autonomy, and supervisory control of uninhabited autonomous systems. Authors may also discuss these concepts as situated within the larger area of human-robot interaction and human-robot collaboration in the industrial sense.

In this paper, we adopt a broad view of hybrid interactions. The balance of agentic AI between the human and machine depends fundamentally on the capabilities of the two systems and the distribution of functions that exist between them. This balance can shift dynamically, as the capabilities of either the human or the machine or the environmental circumstances change. Human-AI collaborations will exist on a spectrum of levels of autonomy, which may necessitate higher levels of oversight, decision authority, and control, or may feature more independent and unsupervised contributions by the AI. At the same time, each hybrid system will have multiple interactive goals, such as effectively supervising the AI, teaching the AI, correcting errors, and learning from the AI. Letting the machine perform more functions may lead to increased workload or a reduced sense of agentic AI. At the same time, human-AI collaboration has the potential to create systems that are able to accomplish more complicated goals than the human or the AI could accomplish in isolation. By operating with integrated reasoning and perception and rich, nuanced interaction, hybrid systems can take advantage of the “best of both worlds.” Successful human-AI collaborations bring synergistic results, adapting gracefully to systematic errors and unanticipated changes in the environment by leveraging the AI in new ways.

2.1. Definition and Scope of Human-AI Collaboration

Human-AI collaboration is a complex concept. At its core, it refers to the interplay between human agents and AI systems across various activity contexts. Layers of the Human-AI collaboration interaction are present in contexts like manufacturing, healthcare, banking, and transportation, where the collaborators are expected to achieve complementary roles or shared objectives. Alongside the dimension of interaction are additional, related aspects such as the manner in which collaboration is concretized. For example, in power plant control rooms, the interaction with AI-complemented systems is a matter of classical Human-Machine Interaction. Here, especially well-studied are aspects such as the “upholstery” of the interface: how information is presented in ways optimized for perception, throughput (for the tasks which are judged important), and emotional stress. A basic attribute of the Human-AI collaboration, “updating the same implicit model of the contexts of the interaction, their own roles, and the means for a successful interaction” remains, however, relevant.

The involvement of a human partner in a collaborative setting is not merely a delightful pairing of two previously separate abilities. At best, the human in the loop has skills, experience, or knowledge of the situation as well as the modes of interaction which allow them to form an effective force multiplier to the AI (or vice versa). At worst, the human will possess imperfect, partial, or out-of-date expertise, lack relevant interpersonal skills (especially in social areas), not manage the interpersonal relationships, or not have suitable conversational strategies. In this regard, “collaboration” includes more than sharing representations or plans. It includes adjusting behavior, effort allocation, and changing priorities in the manner suggested by the other partner in the common goal. In situations where AI interventions tend to be curtailed by ethical or cultural evaluations, for example, instructions from an AI for the use of force, a human actor’s collaboration may result in an action that is not optimal but is ethically, morally, or legally acceptable. The distributed expertise assumption that underpins such collaboration runs up against constrained AI reliability and user stress in a number of well-documented outlier cases. Further, as we discuss in a later summary, collaboration can be considered, in part, as a mechanism to promote agentic AI. In environments where AI autonomy is the norm, occasional collaboration could foster, rather than degrade, user acceptance of AI decisions by providing a perception of control.

2.2. Historical Context and Evolution

The history of developing AI and incorporating it into domains makes various forms of human-AI collaboration possible now. Drawing a comprehensive history of AI is far beyond the scope of this essay, and we refer the reader to several comprehensive historical accounts of AI that are inherently connected to these broader studies as well. Notably, AI has undergone multiple shifts throughout its history and can now be seen to have gone through a few different models of collaboration from both conceptual and technical perspectives. General AI would entail systems that could perform human-level duties and can be compared to systems developed today that are sometimes designed to interact with many different capacities in various spaces. During this time, intelligent agents were proposed, which were systems that embodied AI techniques in software and hardware systems, planning systems in less familiar environments, contract nets which were systems responsible for contracting out tasks to a range of computers, and expert systems which functioned with a great deal of autonomy in separate fields like ensuring the quality of parts in the aerospace industry. Moreover, agents were envisioned that were responsible for checking information conflicts, while very specialized, and there were also envisioned systems that performed a great deal of autonomous tasks in modal logics, including many mathematical tasks.
Questions about what autonomy could entail have been connected to AI ethics and AI as well, with the notion that humans may feel threatened by autonomous AI. It is argued that, while many may believe this has been a main focus in developing AI, the tension this paper discusses is "long-term cooperation among entities, some with very different interests, while still continuing to recognize each other’s autonomy." It is further argued that this autopoietic legal entity could also extend this recognition to autonomous AI in their midst. The question arises, "What algorithm-based developers are trying to produce non-autonomous AI, and to what degree?" However, given the discussion from the history of AI, the need for autonomous AI to perform complex decision-making seems inevitable. This point is also evidenced by the understanding that autonomy includes the ability to stop a process. Thus, for instance, an autonomous vehicle should have the ability to break control to prevent harm, which also aligns with the notion of moral machines. The same feature is a necessary quality in the biologically inspired AI that brings moral and altruistic decisions made by other biologically inspired AI, agents, software systems, or robots.

3. THEORETICAL FRAMEWORKS IN HUMAN-AI COLLABORATION

Over the past decades, human-computer interaction (HCI) and artificial intelligence (AI) research have significantly advanced how humans can collaborate with intelligent software agents. Central to this strand of research is a profound theoretical understanding of the interplay between human and computational behavior. This issue arises, among others, from the combination of human and AI activity on different levels, from intention and action all the way to perception, creation, and coordination of meaning.

Key theoretical frameworks underpin this study. There is a broad range of schools of thought on which theoretical insights relate to the intersection of human and artificial intelligence. Although these frameworks are brought to bear by HCI and artificial intelligence researchers, the objects of their inquiry may also provide insights relevant to data-driven economics. Agentic AI is an important concept when considering theoretical underpinnings of aspects of control; that is, how humans interact in the development and arrangement of the 'superimposition of control' of the composite system as it interacts with an AI component of the same composition. The desideratum in such an active theory of agentic AI is to be able to understand and control the interactive properties in such socio-technical systems to construct a safe and useful system.

As AI technologies become actualized in systems and interact with humans directly, such concepts become critical in their actualization as design constraints in practice. Understanding AI autonomy is important in its implications for collaborative control and influence, human-AI relations and trust, and projecting workflow, for example. "What kind of action patterns can be executed automatically?" Ideally, we will be able to condition more granularly the kinds of communicative exchanges and normative social accountability that make sense in context by controlling liberally autonomous AI behavior. Sociotechnical systems theory informs HCI and AI hybrid interaction theory mostly indirectly by establishing the dynamics of human-technology interaction and practices both informally and institutionally. Given its strong grounding in and reliance on sociological frameworks, the theory of sociotechnical systems brings to bear insight into and operationalizes the investigative aspects of the phenomenon. Humans in such systems, however primitive or otherwise, are today 'working' with the system AI exactly in this phenomenological sense when they are provided with charts of training data.

3.1. Agentic AI and Autonomy in Hybrid Systems

If a robot is regarded as a legal person affords legal rights, then who is to be blamed for the injury or loss occasioned by an independent decision by the robot? According to authors, in hybrid systems people remain free-willed agents while artificial intelligence provides opportunities for machines' learning, deciding, and pro-acting to fulfill their goals. It is clear that the operational dynamics of human and autonomy tease out a dubious distinction, thus requiring a balance for the best harmony.This can cause a lack of trust; disappointment and time wastage in performance of organizational goals. Maintaining this balance involves key factors such as:

- **System Transparency:** Clear AI decision-making processes.
- **Environmental Adaptability:** Handling uncertainty in dynamic contexts.
- **Human Trust:** Building reliability and confidence in AI.
- **Control Dynamics:** The harmonization between decision making structures and definition of human intervention.
- **Time and Context**: Relationship between the level of decision-making AI autonomy and decision urgency and complexity.

Conceptual model from disability studies describes the power and control in human- Al interactions therefore providing direction on how to create effective relationships. In this case, design principles of agentic AI in hybrid systems have the potential of promoting trust, enhanced productivity together with societal alignment, if ethic concerns are well embraced.

Table 1: Key Factors for Balancing Human and AI Autonomy in Hybrid Systems

Aspect	**Description**
System Transparency	The feature lets people understand exactly how and why AI systems produced their recommended decisions.
Environmental Adaptability	Enables AI to handle uncertainty and operate effectively in dynamic and unpredictable contexts.
Human Trust	This method ensures users trust their AI systems because these systems work well every time.
Control Dynamics	The algorithm lets organizations control decision-making processes through defined steps and human participation rules.
Time and Context	The research explores how much freedom AI systems should have when making essential and advanced decisions.
Design Principles	The approach develops systems to build trust with staff and customers plus improve organizational performance while addressing ethical challenges.

3.2. Socio-Technical Systems Theory

Theory Socio-Technical Systems theory is concerned with the interplay between social and technical components in systems. This theory argues that social and technical dimensions continually need to interface with each other to assure efficient system operation. According to this theory, systems rely on humans, as well as on technological processes and tools to steer and operate, and need to be concurrently designed and deliberated accordingly. This mutual adaptability is described as stability. Stability is achieved when human actors adapt, that is, modify their intentions, actions, and cognition in ways that are compatible with the pre-given technology. From a socio-technical systems perspective, it is not only necessary to understand technology for human-AI interaction but also embedded AI systems that can deal with basic user input and execute the required activities.

Resistance can stem from anxiety towards new technologies or the new tasks and organizational roles that come with them, but through commitment, transparency, and early involvement of skilled human operators in the design process, the gravity of this can be strengthened. Further, efforts focusing on workarounds will drive effort and awareness on misalignments and point out key issues. Nevertheless, sharing authority over who is better placed to make decisions in the system, giving or receiving support, and miscommunications can result from misconnections of intentions and expectations. In view of this, alignment refers to adjusting machine capabilities to better represent human intentions. Such adjustments can stem from changes in knowledge and goals, training, or problem formulation. Investigations about alignment take place in real-world task settings. The socio-technical systems perspective anticipates these resistances and shows us the wide-ranging domain of corrective strategies on which to draw. This is the sort of problem we plan to address in this paper.

4. CHALLENGES IN BALANCING AGENTIC AI AND AUTONOMY

Managing and allocating both the agentic AI and the remaining human autonomy in a hybrid system is a very difficult thing. The risks such as inherent autonomy, privacy, as well as on data

sharing need to be managed to ensure that every step meets the standard ways of informed consent. In light of the issues presented in this paper, this issue strongly suggests transparency in AI design and the clear demarcation of responsibilities relevant to AI decision making.

The greatest concern is one of managing developing, complex artificial intelligence technologies. Current laws cannot always effectively cope with the problem; thus, the need for flexible and universal laws adopted with regard for cultural standards and the opinion of experts from other fields. If not well managed, the disparity of human decision-making control and artificial intelligence decision-making power triggers skepticism and poor coordination.

Schroeter said that striking this balance also requires knowledge of what AI can do and cannot do and where humans come up short. By incorporating ethical norms into the design of technologies, developing sound principles at the highest level, and engaging in collaborative research, hybrid systems can deliver desirable interaction with living beings targeting high autonomous and agential value.

4.1. Ethical Considerations

While theoretical or conceptual representations of both active agentic AI and autonomous action are valid, they often seem mutually exclusive in practice. This leads to a number of ethical considerations. From a privacy perspective, the more control the AI system has over its decisions, the more personal information and agentic AI it has access to. Similarly, the capability to make its own decisions also makes AI systems accountable for those decisions. Ensuring AI systems are unaware of people's sensitive characteristics is fundamental to eliminating bias. The more freedom developers allow AI systems to enact their own policies and decision-making approaches, the less control they can have over this attribute. While systems with distorted goals are not necessarily unethical, if a user is operating under the assumption of a different objective, this has the potential to cause more undesirable outcomes than a misunderstanding of functionality alone. Finally, ethical guidelines demand that AI systems must be controllable. Without surrender of a BCI vividly illuminating an AI system's inputs, the question then becomes: to what extent should an AI system make its decision-making processes transparent to its human collaborators? In a practical sense, it is extremely difficult to pre-determine the distribution of skills. The problem of conflicting values, therefore, seems to emerge from competition rather than cooperation. The capacity for ethical outcomes can be impaired or improved on two levels: firstly, by the operating policy of the system, its 'morality'; and secondly, by the fundamental goals or directives given to the system during its development. From this perspective, it is possible to identify the potential for a spectrum of unintentional ethical consequences. While systems designed for pro-social intent do not, by default, ignore the freedom of human agents, there are potential unconscious ethical conflicts at the point of system development to be considered. Primarily, developers need to establish whether ethics is a relevant consideration in the design of their system, and which ethical framework they might adhere to or adopt. They must also determine the extent to which they are willing to comply with these ethical considerations and, if they do, commit to subsequent revisions of their system to adhere to and remain compatible with an ethical structure. Social implications and side effects must be identified, and AI developers and organizations should design their systems in a way that reflects such considerations.

4.2. Legal Implications

The legal framework has to answer the question of who has to bear the responsibility in cases where a decision by the AI system underlying the HAI may have negative consequences, particularly if these negative consequences are a result of a division of labor between the human

operator and the HAI. As AI systems are based on programming processes that in most parts are beyond the individual's control, the question arises whether an AI system can create a situation where neither the person in the loop nor the operator is responsible, but where the outcome is purely bad luck. While legal reasoning might object to the delimitation of responsibility in cases where the programming process is unknown or where unforeseeable events or interacting systems result in unintentionally harming somebody, the extent to which the AI behind the HAI can lead to harmful outcomes by operating as intended is still inadequate. Disputability of AI decision-making is aggravated by the fact that the development stage of AI systems is characterized at least by their self-learning algorithms and high sensitivity to input data, meaning that the outcome of the decision-making process could vary although the subsequent steps taken by the machine may be the same.

Further implementation of Human-Artificial Intelligence interaction approaches has to cover the existing legal frameworks. If data is stored and processed, the legal framework has to respect data protection regulations, and if learned or inferred knowledge has to be handled adequately within the European context, which respects fundamental human rights, particularly privacy. Despite some rather generic legal frameworks, the level of data protection differs widely between countries. But not only do the rights differ, the extent as well as the adaptation shape the existing legal landscape and finally influence scenario planning. Since AI technologies and their entities are assumed to be global and their use is not limited to a certain state, the application of laws should tend towards the higher standard and their universal principles. Yet, the question of to what extent, with the existing patchwork of laws ruling AI systems on a national and international level, one may rely on profound, constructive, and cooperative legislation governing scenarios has to be answered critically. It is necessary to strengthen dialogue between legal, technical, and ethical experts to prepare the ground for corresponding legislation to come. Moreover, uncertainty about how laws and rights are interpreted can contribute to a compromised trust in AI-based collaborative systems. The lower the trust in systems, the more reduced the goal-oriented collaboration of humans with AI becomes.

5. BEST PRACTICES AND GUIDELINES

When designing a human-AI collaboration, it is important that the development is guided by its users' needs and the specialist tasks they seek to solve. General design principles should reflect the purpose of the hybrid system, whether the aim is to increase the role and decision-making of professionals, reduce documentation loads, work more efficiently, or improve the performance of a process or system. Most of these best practices skew towards open-loop AI systems, and research will be needed to understand how to balance these principles when the user is part of the control loop and the task is shared between human and AI.

To achieve intuitive human-AI collaboration, it is important that the use of both agents is intuitive enough for the human end-users, and that the design of the AI agent facilitates the use of both agents together. Both control systems and humans and layered interpretations of how to best engage and problem-solve exist in differentiated literatures. Best-practice social research suggests that engineers have clear principles and human-centered processes, act as facilitators, treat lay knowledge as a resource, and require clear governance, support, and training. Ongoing education of developers and end-users of AI systems is vital for coordination and cooperation with hybrid human-AI systems. This includes training that is part of embedded education of professions and continual professional development and training of practitioners.

1) User-Centered Design: Systems where the human is part of an AI control loop require specific front-end user-centered design to ensure users are empowered to be able to moderate system decisions. 2) Human-AI Hybrid Systems Training: Ongoing educational engagement with groups

of many different professionals focuses on the combined AI and human approach to solve challenging real-world scenarios and dilemmas, and allows trainers to critically examine how new AI technology is used to inform and/or make decisions. This mode of training can be used at any scale at which people are using or designing systems that comprise AI-processed data or advice. 3) Frameworks for End-User and Public Engagement: An established model exists of how to meaningfully engage professionals and others in codesign of data-driven AI advice systems. The results of this engagement include experiences, knowledge, and opinions about the performance of AI Face Decision Support in their local operational contexts. Initial phases of research involved engagement with front-line medical, nursing, and policy end-users of current face decision support pathways, as well as public groups that came into contact with digital facial analysis products. 4) Best Practices for Technological Organizations Involving the User in Best Practice Development: Best practices for including users in the development of information technology have been developed and funded. Organizational participants involved in codesign appreciate being part of the codesign intervention and report benefits at all levels of the case organizations, including personal career advancement, improved project outcomes, and organizational objectives. 5) Human-Like AI Assistance Systems: Advances in computing and AI are allowing for new forms of collaborations where AI-human collaboration is increasingly blended in shared cognition models of hybrid systems. Human-in-the-Loop AI Systems: Provides an overview of the use of HIL AI systems to date and their significance. Describes the potential societal, economic, cultural, environmental, and health benefits of HIL AI systems, as well as any potential adverse or negative consequences for these systems.

5.1. Design Principles for Hybrid Systems

Hybrid systems blend human and artificial intelligence. There are several fundamental design principles critical for hybrid systems. First, the system must be designed considering the user from the outset of the process; this is called human-centered design or user-oriented design. This involves user research, understanding the user’s capabilities and limitations, iterative user testing, and the inclusion of the user experience and wider ethical issues within the design process. Collaborative approaches often extend human-centered design to human-AI collaboration, specifically with a particular emphasis on ensuring the human retains control. Second, agentic AI, or control, over AI systems is an important aspect of hybrid systems where human control is shared with an AI system. A range of design guidelines, recommendations, and methodologies are being developed for creating collaborative hybrids and other kinds of hybrid systems. Iterative user testing is frequently mentioned in the design of these systems, one of the principles used in this work too. Setting clear objectives and designing systems with features that match the skills and knowledge of the user trying to achieve those objectives is important to support the development of human-AI collaboration. It makes use of human-human collaboration as an analogue and a basis to represent and specify the principles of human-AI collaboration. Although the principles of HCI are well developed, the human-AI aspects are still in development. Transparency, in terms of the operation and roles of AI in the system, is seen to be an important principle of collaborative hybrids.

Transparency can also support the establishment of empathy, trust, and social acceptability; all of which can be important in human-AI collaboration. In relation to utility, hybrids must satisfy users' needs and desires. The principles of inclusivity and diversity in design are important. There are principles and guidelines within HCI to guide this process for user interfaces and other technologies. There is potential for adversarial attacks, with humans struggling to interpret AI outputs. Assistive AI and robots also need to be accessible for people with disabilities. An inclusive design approach to robots was proposed, and much of it may also apply to AI systems. In relation to ethics, the principles of transparency have implications for privacy, responsibility, and accountability. Transparency has ethical implications, as a lack of transparency can lead to a

lack of safety and possible entrapment that restricts autonomy. It is important to mitigate having humans trust that the AI has properties it does not have and to avoid humans having excessive trust in a hybrid system. The involvement of users in the design process, including consultation, co-design, and end-user participation, is seen to be important in HCI. This includes the process of user testing of systems following iterative design.

Table 2: The Design Principles For Hybrid System

Principle	**Description**
Human-Centered Design	It centers on bringing users into product design work by studying them and testing with them while recognizing their strengths and adaptabilities.
Iterative User Testing	Users receive ongoing system tests to help perfect features and keep them aligned with system requirements.
Transparency	Sets out why defining AI responsibilities and systems creates trust, empathy and makes AI more acceptable to society.
Collaboration Models	Uses collaboration between humans as its base to develop human-AI interaction rules.

5.2. Training and Education Strategies

In order to maximize effective and successful outcomes in the diverse conditions of human-AI partnerships, training needs to have multiple components and be tailored to the diversity of potential users. Training needs to be structured with clarity and should be scheduled to minimize any disruptions to work. However, the current fallibility and limitations of AI mean that the range of capabilities and knowledge required by the users is extensive. Training for AI required by end users must be multi-level and account for a range of roles that might exist, as well as the varying current capabilities to work with AI from potential users. Given the multidimensionality, the training approaches and structures for current technologies should be designed to anticipate what might eventually become a more complex training regime. Practical, experiential learning should be prioritized where possible. Other strands of work in education and learning may also be useful to this review. Multi-disciplinary learning, where both technical and human/social researchers in the field can learn together, would usefully emphasize the application of design principles that are relevant in a range of different knowledge domains. This aspect of designing training structures that are beneficial across different fields is also relevant to the subsequent discussion about collaborative spaces and community building. Work in psychology also highlights how essential it may be for those seeking understanding of collaboration with AI to develop creativity, exploration, and playfulness, which having a safe and nurturing community might facilitate. We return to this point in our section on community building. Observational and field-based research would be invaluable to this review in order to ground the work in real-life practices and document the essentials of collaborative practice. Almost as a conclusion here, we ask what other practice, expertise, and learning dimensions we have missed that are relevant to the integration of AI. This question resonates with the final section that discusses our practice-based focus for this area.

6. CASE STUDIES AND APPLICATIONS

In this section, we discuss a number of case studies. These case studies are aimed at showing the process in which human-AI interaction has been developed in the real world, the results and lessons from their implementation, the relevant practices and challenges that they highlight, and the process of ethics monitoring and evaluation that was conducted as these were developed. Each case study explores an application in a different domain and demonstrates both the challenges and the potential of integrating human users into AI systems in practice. By examining

the processes of human-AI interaction involved in projects across these different sectors, we can extract likely applications and begin to identify some of the methodologies that apply across such diverse practices.

The first two applications described here are situated in automation systems: the former in an air traffic control simulation and the latter in the context of controlling a quantum device. In both cases, researchers are partnering with domain experts and designing interaction with autonomy into interfaces. The third case is based in pedagogy; it is an application of machine learning to detect learners' attention during an online course. Two examples of systems designed to detect their users' errors are given; the latter acts as a tutor. Two further case studies were conducted in the medical domain. In the first, machine learning was used to develop a risk matrix that was used to inform the emergency teams responding to issues raised via Medical Emergency Teams. The second offers an alternative approach to human-in-the-loop processes, where the automated partner is used to generate human training data by providing an initial hypothesis, which experts could then apply to increase the detection rates for endoscopy errors. The final four examples are varied, with the first describing an application of machine learning to financial data and a system developed to support steering an L3 autonomous car.

The following four system developments have taken place in the UK. This includes the description of a machine learning system in support of the Research Excellence Framework Panel. Following an introduction to the segmentation recognition benchmark used by the study, this section describes the sentiment analysis-oriented case study. This project applies recent advances in vehicle and pedestrian detection to enable autonomous vehicles to differentiate between a vehicle that is stopped and one that is about to pull out. This may support, for example, autonomous vehicles to exit junctions without colliding with other traffic. The project develops a visualization of derived point clouds for the Oxford Robotic Car by experimenting with three state-of-the-art semantic segmentation algorithms. The visualization allows users to monitor the autonomous car's perception of its environment as accurately as state-of-the-art point-based segmentation algorithms. In this case, the methods underwent validation by both human and machine assessment.

6.1. Healthcare Sector

Health: The integration of AI technologies in the healthcare domain is a promising opportunity to enhance service and healthcare delivery. This integration allows clinicians and medical personnel to enhance and expand their current capabilities. The healthcare sector is undergoing a digital transformation, wherein many processes have to be re-engineered and rethought in order to leverage AI systems. The use cases for AI in healthcare range from non-clinical to clinical. Examples of non-clinical use cases include administrative support such as patient satisfaction surveys and customer relationship management. Clinical use cases range from diagnostic to predictive and personalized capabilities. In the highly conservative domain of healthcare, collaboration between AI and humans is ethically and morally significant. As a result, interdisciplinary research through collaboration among technologists, ethicists, patients, providers, and payers is essential in defining training and use strategies of hybrid systems. The hybrid system must maintain and build patient and provider trust in medical AI systems, as well as confidence. Moreover, technical development needs to be informed by genomics, data privacy, ownership, and data protected by laws. The system must guarantee ethical and strategic methodologies for deciding what can and what should be communicated to patients and when. The stakeholders must develop AI systems that are trained on a wealth of protected health information and implementations that work and that are in progress present lessons about building the capacity of healthcare systems to accommodate the enormous flow of digital data being shared.

6.2. Autonomous Vehicles

AI systems have transformed mobility and transportation, particularly gaining a lot of attention in the area of autonomous vehicles (AV), with the potential to adapt the interaction between AI, digital infrastructure, and humans. One of the key values of AI to AVs is their potential to reduce human operator error; human error is a leading cause of road accidents and can account for a significant percentage of such accidents. Different AI systems, such as rule-based and machine learning models, have incorporated AI technologies in their control strategies, ranging from simple dynamic system modeling to perception, reasoning, and planning from environmental inputs and actuation commands. The representation of humans in the AV can be either as a monitoring and intervening presence in the AV navigation or in a shared-control capacity. For example, human drivers of an AI-based automated shuttle servicing city streets are in charge of monitoring the environment and condition of the vehicle, supervising the door opening and closing, and taking control of the vehicle's navigation if necessary, using a steering wheel and/or a touchscreen.

The role of the human operator in current-day AVs ranges from blind de facto operators of AI systems to likely future drivers. What is interesting is that the trend seems to be moving the function of human operators from making decisions—whether it's steering the vehicle or deciding not to take action in emergency scenarios—to a purely monitoring role. Just like air traffic control operators, train and metro operators, and pilots of military drones have been doing for decades, an operator's core function in an AI system might be that of a supervisor, monitoring the performance of the system and intervening in case of malfunctions and other abnormal conditions. Deployers should put in place systems to account for accidents and near-accidents, possibly establishing an investigation process involving the inspection of the vehicle and data logging of a relevant period of AV operation before the accident, including data and self-reflection of the human supervisor operator. We assume that any regulatory requirements would be guided by national or even international standards, operational rules, and other technical guides. Once again, training needs to be imparted to turn the vessel operator into a system supervisor agent.

7. FUTURE DIRECTIONS AND EMERGING TRENDS

This chapter reviews some of the technologies and possible design futures regarding human-AI hybrid systems. A variety of rapidly evolving technologies offer opportunities to increasingly support human sense-making, deliberation, and efficiency in information and communication as never before. In computer science, developments in the field of machine learning and deep learning have already started to provide new ways to outperform human competence in certain tasks related to analysis and pattern recognition. Moreover, a range of AI technologies have been enabling natural language understanding and generation capabilities that have gone beyond conventional chat systems.

Moreover, the same technologies can also align with the aspirations of symbiotic action, where AI can be integrated alongside human skills and attributes to amplify and augment human performance. Robotics research has been much interested in how increasingly autonomous robotic systems can be developed to engage in more sophisticated, resilient, real-world collaboration with human beings, e.g., within logistic scenarios or repair tasks, or perhaps one day, as co-present team members. Any or all of these developments will have profound consequences for the human workforce and the kinds of job roles and associated skills that will be required in the future. Like other disruptive technologies, AI could lead to shifts in the kinds of employment opportunities that are available, perhaps liberating workers from some repetitive

tasks and dull routines or creating new kinds of work around AI development, or, conversely, eradicating traditional choices within sectors. It is the societal capabilities to flexibly adapt and reskill that will determine our capacity to manage such change effectively. Indeed, proactive engagement strategies will be critical in helping us to deal with the practical consequences of these convergent trends. New kinds of governance frameworks and ethical considerations may also need to be developed. Still, the AI community must value itself by manifesting creative and thoughtful reflection about these emerging near-term and long-term possibilities.

7.1. Advancements in AI Technologies

The advancements in different AI technologies are becoming a beneficial factor in realizing the potential of human-AI collaboration. Primarily, the advancements in machine learning techniques, such as deep learning and recurrent neural networks, that can process complex multimodal data, such as images, videos, texts, speech, and data from a variety of sensors, are instrumental in transforming the state of AI. Unlike traditional statistical models, which require handcrafted features and can process limited amounts of pre-processed data, AI models can take advantage of unsupervised feature learning on raw data in a scalable and parallelizable manner. The advent of controllers, connectionist representations, and algorithms for learning and decision processes, such as deep and generative learning, lifelong learning, hybrid optimization, learning to learn methods, reinforcement learning, and distributed learning and reasoning, leads to exploration scenarios. The availability of big data systems and distributed computing enables fast AI training in a synchronized and asynchronous manner, facilitating real-time data processing for quicker results. The availability of advanced algorithms and software accelerators for hardware enables energy-efficient computation.

The decision-making and object interaction quality are steadily improving due to these AI and robotics advancements. Especially, AI being data-driven, it can be modeled as a function of a variety of data inputs from different sensors and databases. If the data input sources are reliable, capturing potential diversity of viewpoints, and are inclusive, such AI models can learn data respecting both formal feelings and informal feelings that ignore or are unaware of some orthogonal relationships in humans. More importantly, AI algorithms are being designed to be 'memorable' so that they collate user preferences and can recommend alternative placements with probabilistic rankings when there are many other people that correspond to the desired options. Unlike traditional or even the state-of-the-art recommender systems, the underlying collaborative or competitive learning paradigms deploy group and social conspiracies and biases towards champions or adversaries that protect the social subgroups or sensitive identities and the vulnerability of who among the individual participants are adopters, defectors, or agitators. All these future-looking approaches are still challenging; current ones are based on closed-world assumptions and also exhibit group biases. A closer look at AI explainability, agreement arbitration, or designing for norm-adopting models is still on the horizon to be done from a requirement level. Human-AI collaboration is essential for the verification and validation of an AI model en route center and can be no less than an AI as no two humans are alike. AI has the potential to be retrained and centralized for either an industry leader or common denominator. AI can be inclusive or exclusive in a way that seems very uncanny at first glance, so it is advisable that in the design of AI technology all stakeholders come together for ethical, responsible, and sound AI making in harmony. AI researchers will attempt to bypass the guards of norm- or intent-respecting, but a lot of work needs to be done in collaborative cross-validation of AI and AI-capable human-in-the-loop empowerment to work together in the future.

7.2. Impact on Workforce Dynamics

The integration of AI has attracted considerable attention not only in AI technologies themselves but also among user communities in terms of its impact on our daily lives and job opportunities. Around half of the participating companies expected to change their main job roles in their industries by the year 2025, partially due to the integration of AI technologies. AI has reshaped and will continue reshaping the human workforce dynamics mainly in two aspects. Firstly, it will directly change the amount of workload dedicated to the employees. AI technologies are complementing, rather than substituting, human skills in these switched job roles, which leads to more workforce increases. Secondly, it indirectly changes job skills and occupation types for future job seekers and holders. Thus, companies should recognize and continually correct job content so that the problem of mismatched skills available in the market can be resolved. This can only be done through continuous education, training, skilling, and reskilling initiatives, such as through alliances with educational institutions, policymakers, and labor unions.

Some important studies have emphasized the further workforce opportunities resulting from international span. It was reported that only 10% of the employee analysis data were occupations related to science, mathematics, engineering, and technology, while the other 90% of the top-natured occupations require minimal digital skills of the office worker in areas such as text processing, green tasks, and more multimedia job areas. Broadly, a series of worldwide companies have also reported on the changing job roles driven by the influence of AI, where the estimation of the market demand for robotics and AI has significantly raised with lower demand in developed countries and higher demand in leading global economies not until 2030. Consequently, AI was expected to adopt more jobs or occupations in developed countries than in developing countries in the same period. However, while AI and robots in assistance should receive a decent amount of attention in several countries' efforts, ethical and normative implications of the demographic shift should be considered. For instance, transitions or losses in jobs may be a source of hardship for families and societies. This transformative process is predicted to bring about some reconstruction of policies in countries, such as a shift in social policy debates about employment to a more comprehensive economic debate because the market for goods and services is now no longer what we have been used to considering it to be.

8. CONCLUSION

In closing, the ability of AI systems to produce and innovate new technologies and processes is critical to the advancement of emerging fields across the physical and social sciences. Moreover, as AI systems grow in capability and flexibility, we risk facing an existential threat to our own livelihoods as humans if innovative new designs are divorced from human-relevant constraints. We have laid the groundwork for delineating trade-offs between pushing the boundaries of high-fidelity task performance and enabling collaboration and interaction with humans. We have proposed a more comprehensive definition of agentic AI to guide the design of AI systems that watch or work with people, as well as diverse application scenarios that highlight the range of hybrid system abilities that must be balanced in order for people and AI to collaborate effectively. We use these application scenarios to identify a set of trade-offs to be balanced when designing AI systems to work with people in these contexts. Finally, we focus on the notion of AI autonomy and what it means for AI to be autonomous, as well as for the AI community and society to be prepared for increasing levels of AI autonomy. Taken together, this provides a broad perspective on many of the challenges of human-AI collaboration that arise when AI systems do work that people care about, from autonomous vehicles to AI assistive technologies and automated programming systems. Our hope is that this work provides a cohesive foundation for thinking about how to balance the need for AI systems to achieve high-fidelity performance with the

desire to produce AI systems that can work well with people. Just as autonomy requires that people are prepared to provide particular types of support for AI systems, AI should be prepared to provide appropriate types of support for people.

REFERENCES

[1] A Chen, M Xiang, M Wang, Y Lu - Information Technology & People, 2023 - emerald.com. Harmony in intelligent hybrid teams: the influence of the intellectual ability of artificial intelligence on human members' reactions. https://doi.org/10.1108/ITP-01-2022-0059

[2] S Caldwell, P Sweetser, N O'donnell… - ACM Transactions on …, 2022 - dl.acm.org. An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. https://doi.org/10.1145/3514257

[3] MMM Peeters, J van Diggelen, K Van Den Bosch… - AI & society, 2021 - Springer. Hybrid collective intelligence in a human–AI society. https://doi.org/10.1007/s00146-020-01005-y

[4] I Munyaka, Z Ashktorab, C Dugan, J Johnson… - Proceedings of the ACM …, 2023 - dl.acm.org. Decision making strategies and team efficacy in human-AI teams. https://doi.org/10.1145/3579476

[5] R Zhang, NJ McNeese, G Freeman… - Proceedings of the ACM …, 2021 - dl.acm.org. " An ideal human" expectations of AI teammates in human-AI teaming. https://doi.org/10.1145/3432945

[6] G Zhang, A Raina, J Cagan, C McComb - Design Studies, 2021 - Elsevier. A cautionary tale about the impact of AI on human design teams. https://doi.org/10.1016/j.destud.2021.100990

[7] CR Sauer, P Burggräf - Production Engineering, 2024 - Springer. Hybrid intelligence–systematic approach and framework to determine the level of Human-AI collaboration for production management use cases. https://doi.org/10.1007/s11740-024-01326-7

AUTHORS

Gaurav Samdani has over 18 years of experience leading automation, AI and data analytics teams

Ganesh Viswanathan is an accomplished technology leader with expertise in AI, cloud engineering, and intelligent automation. Currently serving as the AVP - Senior Principal AI Engineer at MetLife, Ganesh plays a pivotal role in driving the company's technology vision, modernizing infrastructure, and leading AI initiatives like intelligent document processing and mainframe modernization. With a strong background in test automation and cloud solutions, Ganesh previously contributed significantly at Ally Financials. Beyond his professional achievements, he is a dedicated family man, balancing his career with parenting two school-aged children. Ganesh is also data science enthusiast with master in data science and business analytics and is passionate about learning and exploring cutting-edge technologies like Ai agents and GitHub Copilot.

Abirami Dasu Jegadeesh is a Energetic, motivated, and diligent professional with 18years 2month(s) Software Architecture & Development experience in Web and Mobile based applications with good communication & interpersonal skills, a highly motivated & productive individual. ▪ Adroit at learning new concepts quickly and communicating ideas effectively. ▪ Dedicated and highly determined to achieve personal goals as well as the organizational goals. ▪ Ability to work both independently and as team to achieve results in stated standards.