Augmenting Human Intelligence - the IBM Point of View

IBM

Attribution

With gratitude to the AI Ethics Board workstream's executive sponsors, Francesca Rossi, Jesus Mantas, Glenn Finch, and Steven Astorino, and the contributions of workstream members Alina Glaubitz, Michael Muller, Christian Busse, Isabelle De Brabanter, Jamie VanDodick, Sophia Greulich, Angel Montesdeoca, John Richards, Werner Geyer, Jennifer Kirkwood, Jochen Friedrich, Thomas Baudel, and Michael Hind.

Table of contents

Executive Summary

Artificial Intelligence (AI) has the potential to build upon the capabilities, experiences, and insights of human beings, transforming how we approach various tasks and challenges by supplementing human-like abilities. The process of augmenting human intelligence refers to the use of AI to enhance human intelligence, rather than operating independently of, or replacing it. This use of AI is designed to include and balance human oversight across the AI lifecycle. Thomas Watson Jr., IBM's 2nd President, once said, "Our machines should be nothing more than tools for extending the powers of the human beings who use them," and we continue to stand by this point of view today.

IBM's Principles for Trust and Transparency delineate the ways in which IBM embeds trust into its technology solutions. The first of these principles is: the purpose of AI is to augment human intelligence. All AI systems developed and deployed by IBM are designed to augment, not replace, human intelligence. IBM achieves this by designing such systems to enhance and extend human capability and potential while simultaneously enabling humans to take on more value-adding tasks. Accordingly, AI should improve job performance and augment individual as well as business performance. This implies that AI systems are to act as support functions to enhance human intelligence, rather than operate as completely separate, autonomous entities.

AI should be designed to include and balance human oversight across the AI lifecycle: when operations are carried out by AI, humans, or humans augmented by AI. This further supports IBM's ability to embed ethical governance into developing, deploying, and building AI systems that help clients succeed.

Historical Context

AI-driven intelligence augmentation can bolster human potential by empowering individuals to navigate increasingly intricate and competitive organizational environments. This empowerment has the potential to benefit individuals and society at large.

During the 1940s and 1950s, visionaries such as Alan Turing, John von Neumann, and Norbert Wiener laid the groundwork for AI and intelligence augmentation, enthralled by the notion of machines amplifying human intelligence. In 1945, Vannevar Bush proposed the Memex — a hypothetical apparatus intended to categorize and index information. Though never realized, this proto-search engine signaled the prospective convergence of human intellect and technology.

Subsequently, J.C.R. Licklider's seminal work on human-computer symbiosis during the 1960s motivated researchers to conceptualize a future where humans and computers cooperated to address intricate challenges. Licklider's outlook portrayed a future in which machines reinforced human cognition. Pivotal advancements at the Augmentation Research Center, led by Douglas Engelbart, resulted in the invention of the computer mouse and the development of interactive computing systems. These innovations augmented mental capabilities by fostering more efficient and intuitive digital environments.

During the 1980s, Marvin Minsky investigated the nature of human intelligence and cognition, establishing the basis for AI systems that emulated human-like cognitive processes. His work afforded insights into the realm of intricate mathematics and facilitated the development of AI tools to assist humans in comprehending complex cognitive structures.

Expert systems and intelligent agents emerged in the 1980s and 1990s, supporting decision makers across various sectors. These AI-facilitated advisors paved the way for the advanced decision-making tools businesses rely upon today.

In the 21st century, machine and deep-learning technologies welcomed a new era of artificial intelligence in which AI systems excel in perception-related tasks, such as natural language processing (NLP), computer vision, and speech recognition. Intelligence augmentation driven by these AI capabilities has revolutionized and improved human problem-solving approaches and decision making.

Today, generative AI techniques provide AI with the ability to generate content, besides interpreting it. Although we are just starting to explore the power of these new capabilities, generative AI is increasingly used to enhance humans' informed decision making ability, creativity, coding performance, and businesses' handling of complex decision making scenarios. For more information on IBM's point of view on the ethics of generative AI and Foundation Models, read our POV.

As progress continues, the synergy between human intelligence and AI will evolve further, molding the business landscape and unveiling untapped potential. The future of intelligence augmentation offers boundless possibilities, reminiscent of the modest calculator that revolutionized humanity's relationship with mathematics.

Current Climate

Economic Implications of Augmented Human Intelligence

The economic implications of AI-driven augmented human intelligence are multifaceted and lead to advantages and disadvantages in all layers of society and organizations.

Augmenting human intelligence can increase productivity as AI-assisted workers become more efficient and accurate in their tasks. This productivity growth can translate into higher organizational performance and overall economic growth, driving innovation and generating new industries. However, the extent to which productivity gains will be realized depends on how effectively AI systems are integrated into existing work practices and how well the AI system and workforce adapt to each other.

The rise of augmented human intelligence may result in shifts within the labor market. This shift may result in transforming employees' careers with an opportunity to upskill into new and different career paths, potentially enabling the AI system to handle less specialized or outdated skills. As AI systems augment routine tasks, workers will be able to focus on more complex tasks that require critical thinking and subject matter expertise. In addition, the increased need for AI specialists and data scientists presents new employment opportunities for those with the required subject matter expertise.

One of the challenges of an AI-driven economy is the further stratification of income inequality, where benefits could be concentrated among a few individuals and thus exacerbate existing disparities. To address this concern, employers can look for opportunities to provide comprehensive employee training and reskilling programs, thereby fostering a more inclusive and diverse workforce that can share in the advantages brought about by these innovations.

Potential Concerns around Automation

IBM Research has studied human perceptions of AI, revealing that humans may exhibit biases in response to AI. Humans have been shown to exhibit automation bias: the tendency of a human to always follow the recommendation of an algorithm instead of their reasoning or intuition, also referred to as 'over-reliance'. Humans have also been shown to exhibit resistance: the human tendency to decline the use of AI or override AI recommendations, also referred to as 'refusal'. These biases inform humans' responses to AI systems and the landscape for human-AI collaboration.

IBM Research has studied human perceptions of AI, revealing that humans may exhibit biases in response to AI. These biases inform humans' responses to AI systems and the landscape for human-AI collaboration.

Although some of the currently available AI systems may generate content similar to what a human being can produce (such as images generated by DALL-E or text generated by ChatGPT), they are not to be treated as human beings. Since

these systems have no conscience, are not sentient, and do not create their own goals, they are not intelligent according to the notion of human intelligence.

A mature approach to human augmentation should include user-centered information about when AI can be trusted and when it cannot. Trust is likely multidimensional: we might trust individual computations but distrust those computations in certain contexts or for certain people. For example, minoritized or marginalized people might require evidence that an AI system operates fairly for people like themselves. In certain situations, individuals or communities may believe that an AI system would produce biased outcomes for them and adjust what data they are willing to share with the AI system, or how they will critically examine its outcomes. Like usability, trust is not a property of a technology. Trust is an attribute of the relationship between technology and an individual user or a community of users. This trust may be different for different users or communities of users.

Furthermore, within workplaces, the implementation of AI augmentation for one individual within an organization could inadvertently lead to negative consequences for an adjacent individual. This situation arises when the tasks of the second individual are rendered obsolete or significantly diminished, as their activities are replaced or overshadowed by the augmented capabilities of the first person. The phenomenon, Shift-right Intelligence-augmentation (SRIA), is critical in understanding the broader implications of AI implementation within organizations. It highlights the potential challenges and complexities associated with workforce dynamics, employee roles, and skill requirements. This concept emphasizes the importance of careful planning, comprehensive employee training, and potential reskilling programs.

Standards and Regulatory Perspectives on Human Oversight

Standards organizations have been working towards including aspects of human oversight into the scope of AI standardization work. In the leading international committee for AI standardization, ISO/IEC JTC 1 SC 42, several standardization projects are underway which focus on, or take human oversight into consideration. These include ISO/IEC TS 8200 (controllability of automated artificial intelligence systems) and the proposed work item ISO/IEC PWI 18966 (oversight of AI systems). In addition, there are several other standards which are related to, or include aspects of human oversight, including ISO/IEC 22989 (AI concepts and terminology) which provide foundational frameworks and terminology for human oversight; ISO/IEC 38507 (AI governance), providing human oversight from a governance perspective; ISO/IEC 23894 (AI risk management) where human oversight is introduced as one of the risk management principles; and ISO/IEC 42001 (AI management systems) where human oversight is included in the controls and implementation guidance for building AI management systems.

Human oversight is also a major requirement under the upcoming EU AI Act. It has been included in the EU Standardization Request issued by the European Commission to the European Standardization Organization CEN-CENELEC where development on a harmonized European standard is starting, within the scope of the work group CEN/CENELEC JTC 21 WG 4 “Foundational and societal aspects.” Given existing collaboration between ISO/IEC and CEN-CENELEC, there will likely be close linkage in standardization work around human oversight in order to maximize synergy and have alignment between international and European standards.

IBM Research

Human-Centered AI

Human attitudes toward AI diverge and may be influenced by individual or collective beliefs in technological fairness; the balance of power between employee and employer; and the balance of power among citizen, community, and government, among other considerations. IBM Research has shown that IBM employees may trust AI to complete certain activities well, but not other activities. We demonstrate these findings through three IBM studies.

Study 1. In an IBM Research study of AutoAI, Dakuo Wang and co-authors studied IBM employees' attitudes toward AI, based on an earlier analysis of the distinct roles and activities in data science teams. Wang et al. found that people in each of their roles thought that their activities were too complex to be automated. However, they tended to believe that other activities could more easily be automated. In the language of trust, the people in that study trusted AI to take over other people's jobs, but not their own. Another interpretation of this outcome is that each person trusted their own human skills to perform their jobs but did not trust automation of their role. However, they were willing for the corporation to use AI to automate other jobs.

This is a microcosm of the broader labor-management issue of who owns the competence to do a job. Management often prefers to hold that competence, and to build it into technologies that can organize the work of lower-skilled, lower-paid workers. The labor force often prefers to hold that competence in workers' individual and collective knowledge, leading to higher skill requirements and higher pay. Numerous studies of invisible work (i.e., invisible to management) and articulation work (i.e., the invisible work that makes the visible work possible) suggest that most jobs require tactical and strategic knowledge that is difficult to capture and difficult to instantiate in technologies. These analyses emphasize the need for using AI to support human skill, autonomy, and accountability, rather than delegating those attributes to an AI system.

Study 2. In a second IBM Research study, Upol Ehsan and colleagues asked two different groups of users (subjects were not IBM employees) to rate the quality of explanations that were narrated by a simulated robot as it went about a task. The two groups were computer science majors (CS group) and people without a CS background (no-CS group). Each group used different information from the explanations, and each group formed different evaluations of the explanations and the robot's performance.

This study showed the need to provide explanations using language and concepts that would be useful for each group of users. This paper also implied that end-users would be more likely to trust AI systems if the documentation and explanations were written so as to be useful for those end-users. In AI documentation and system-generated explanations, one size does not necessarily fit all. It is necessary to consider the needs and vocabularies of the end-users.

This study showed the need to provide explanations using language and concepts that would be useful for each group of users. This paper also implied that end-users would be more likely to trust AI systems if the documentation and explanations were written so as to be useful for those end-users. In AI documentation and system-generated explanations, one size does not necessarily fit all. It is necessary to consider the needs and vocabularies of the end-users.

Study 3. In a series of experiments with human-AI co-creativity, Michael Muller and colleagues used a conversational UI linked to a LLM to explore analogy-based conceptual design. In the first two experiments, the human proposed a design metaphor, referred to as a "framing" by AI creativity researchers. The human then guided the AI to fill in design details within this metaphor. In a third experiment (in the same paper), the human asked the AI to propose a framing, and then the human asked

the AI to fill in the AI-proposed metaphor. Finally, in a fourth experiment, the human asked the AI for a metaphor, and then rejected that metaphor. The human asked the AI for a new metaphor, which is a crucial AI creativity move called "reframing." The human then guided the AI to fill in design details within the reframed metaphor.

This study demonstrated the power of a well-tuned conversational UI. It also demonstrated that co-creativity can occur in the interactional space between humans and AI, with the human firmly in control of both the strategic creative goals and the conversation tactics to achieve those goals.

Synthesis. The broader research literature has reported that AI systems have the potential to produce biased outcomes for minoritized and marginalized populations. To address fairness, researchers have developed tools, such as IBM's AI Fairness 360, to test for potential adverse impacts on minority or marginalized peoples or other groups on the basis of their identity or status. However, industry adoption of those tools is lagging – particularly in the intense competition for generative AI mindshare and market share – and more work is needed.

IBM Research has summarized multiple aspects of these risks of harm in our Foundation Models for Designers guidelines and an award-winning CHI 2022 paper. Unsurprisingly, people who have received benefits from AI systems may tend to trust them – or over-trust them – while people who have been harmed by AI systems may tend not to trust them – or under-trust them. As mentioned previously, these positions have also been referred to as over-reliance (automation bias) and under-reliance (resistance).

These issues of relative benefit, risk, and harm in relation to human-centered AI are complicated because the benefits, risks, and harms are not inherent attributes of the AI system. Rather, they emerge in configurations of humans and institutions. Similarly to the results of the Ehsan et al. study mentioned above, the outcomes may be different for different groups of "users," although it is important to consider that some of the users do not engage with these systems voluntarily (e.g., criminal suspects, bank loan applicants, families facing loss of child custody).

In effect, there are at least two very different groups of stakeholders in these decisions: the individual and a larger organization or society. For example:

- A bank wants to minimize loan defaults, so it will tend to make conservative decisions about granting loans. That conservatism may have an adverse impact on minoritized and marginalized people. Consequently, a person classified as a poor credit risk may not receive a loan and therefore may not have an opportunity to improve their credit risk profile.

IBM Research's Humble-AI project has examined these kinds of tensions. In the above example, there are very different values for the individual as contrasted with the institution. The cost of a false-positive outcome (e.g., the algorithm classifies the applicant as a risk for loan repayment) is relatively low for institutions but is relatively high for individuals. In general, the risk threshold is set by the institution, and it thereby reflects the perceived interests of the institution. Humble-AI describes approaches that can be used to reconsider the risks and the very different costs of false-positive and false-negative outcomes, and to readjust predictive factors and risk thresholds toward greater fairness and accountability.

Human-AI Relationships

Within research around human-AI collaboration, one popular model describes a trade-off between humans and AI: if the human has more autonomy, then the AI system has less autonomy, and vice versa. However, Shneiderman's 2022 Human-Centered AI model argues that AI autonomy and human autonomy may be independent of each other. Earlier work in Mixed Initiative Creative Interfaces and IBM Research's work in Mixed Initiative Generative AI Interfaces explores how autonomy and control may shift from moment-to-moment for highly-interactive work completed by humans in collaboration with AI. For more complex tasks – especially business applications – this flexible division of autonomy seems to be a requirement.

In addition to the division of autonomy between a human and an AI system, it is important to consider the sequence of roles in a developing relationship between a human and an AI system. As mentioned previously, human-AI collaboration can be achieved through a progression of roles for an AI system supporting a human. For example, a new employee may need assistance from an AI system to learn a new job and develop the related skills for the role. The AI may act as a tutor, correcting the new employee's mistakes. As the employee gains skills and experience, the AI may transform into an advisor, providing non-binding recommendation services. As the employee learns more, the AI may again transform into a peer (e.g., human-AI peer programming). At an even later stage, the experienced and highly skilled human may prefer to use the AI as an assistant, capable of doing certain routine and monotonous tasks. At this point, the responsibility distribution has reversed, and it's now the human's responsibility to correct the AI system's mistakes. Consequently, human autonomy and accountability increases through the AI's role transitions.

AI Decision Coordination

IBM's latest survey on AI adoption, summarized as part of the Global AI Adoption Index 2022, suggests that despite an uncontested interest towards AI (77% of companies consider using AI for their business), a significantly lower share (35%) have managed to integrate AI into their processes. One of the possible explanations for this gap between interest and adoption of AI is that businesses generally lack objective gauges of AI performances compared to human decision making, and consequently are unsure of how to leverage AI recommendations. Decision makers may be unable to justify their AI implementation choices and thus are hesitant to integrate AI models into their business processes.

AI Decision Coordination

AI Decision Coordination, an innovative IBM solution, enables managers to orchestrate processes involving both algorithmic and human decisions. It leverages empirical evidence of the relative performance of both algorithms and humans, as well as the cost structure of decision making, to provide robust optimized allocation strategies in routine processes such as alarm filtering, fraud detection, and customer support call centers. The methodology implemented by the tool accounts for cognitive biases, such as automation bias, anchoring bias, and decision fatigue, focusing human resources where they will most likely perform best, while reducing repetitive human workloads where they are shown to be unproductive.

With AI Decision Coordination (AIDC), management can determine the best use of decision automation tools with quantified evidence, and provide an explanation for their implementation choices while making the best of their human resources by allocating them to the tasks where they bring the most value. AIDC achieves this by:

- Building relevant performance metrics through an advanced and modular cost-benefit model
- Assessing empirical and relative performance of AI systems, human, and AI-human collaboration in decision making
- Evaluating expected gains and enhanced quality related to the use of AI in target processes

The key output of the AIDC software is a set of robust allocation rules between AI, human and AI-human collaboration decision making that maximizes the performance (and the associated gains) of the targeted process, according to the business metrics and attributes of the use case.

The above graph raises questions around whether under certain circumstances (e.g., when there is no potential adverse impact on humans), it is preferable to have decisions be made solely by AI-driven solutions, as opposed to under all conditions enabling human oversight and decision making. Human decision making could be redirected elsewhere, and the overall performance of the process could considerably increase. Paradoxically, allowing AI only decisions may actually augment human performance in certain instances. This raises the question of how we want to augment human performance, and which metrics we should use to assess this augmentation.

Sample Use Cases

Finance

Sample Use Case 1: Fraudulent Transaction Detection

In Finance, AI can play a pivotal role in fraudulent transaction detection, a significant concern within the financial sector and society more generally. By scrutinizing patterns within transaction data, AI systems can pinpoint anomalies that might suggest fraudulent activity. Each anomaly can be flagged to a human expert for review, enabling human oversight and the human's position as the ultimate decision maker. Such insights can assist banks and other financial institutions in minimizing losses and safeguarding their customers.

Banks lose $3.5 trillion annually to fraud, having severe consequences both on the lives of the bank's clients and its employees. An IBM client, a European bank, deployed AI models to enhance their fraudulent wire detection and rationalize the number of false alerts that their employees had to review. Despite a considerable reduction of false fraud alerts, the bank's clerks were still feeling overwhelmed by non-relevant alerts and were missing significant amounts of fraud. Even though the use of the AI tool relieved employees from fastidious and painful false alerts to review, the use of AI in this particular case was not correlated with an augmentation of human intelligence, as the bank clerks' performance in terms of fraud detection did not improve.

To remediate that, this bank used IBM AI Decision Coordination to:

- Compare the AI models' performance with the actual behaviors of bank clerks
- Translate these behaviors into relevant business terms (time spent vs. actual fraud detected)
- Define which process changes would build upon the AI models' transactional intelligence, employees' relational intelligence and the compliance team's fraud expertise

In practical terms, the analysis showed that:

- Some decisions should be taken by the AI models without requiring systematic human interventions, as in the best of their abilities, the bank's clerks would have very low chances to spot the 0.1% relevant fraud alerts
- Some decisions should involve the bank clerks' intervention to call the client and determine whether the fraud alert should be taken seriously
- Eventually, the alerts where the AI models predicted a very high likelihood of fraud could involve both a review by the banking clerks and the investigation of a team specialized in fraud

This fraud case study clearly illustrates how the use of AI can have various implications in terms of human intelligence augmentation. It shows how this augmentation might come with a reduction of the number of the decisions requiring human intervention, but also highlights the strength of balancing both human and AI expertise.

Sample Use Case 2: Money Laundering

AI enabled fraud detection has also been applied in Anti-Money Laundering Alert Systems. The fight against money laundering and terrorist financing requires financial institutions to be extremely vigilant in their screening of transactions in order to avoid facilitating such transactions, which can result in, among other things, ethical and regulatory violations and hefty fines. As a result, most of them use highly conservative alerting systems, generating up to 95-99% of non-relevant alerts to be reviewed by skilled financial security analysts. Financial Services actors have taken steps to leverage AI models to eliminate a significant part of these false positives.

The investment banking arm of a leading European bank used AI predictive capabilities to predict these false positives. It eventually managed to reduce these non-relevant alerts by 30% while keeping the same level of risk exposure. For the remaining 70% of alerts, it also used AI to classify them according to categories of risk. These categories proved to be useful in prioritizing the financial security team's workload and optimizing the team's efficiency.

However, while the analysts carefully reviewed the risk assessments at first, after a few months of use they started to blindly follow the AI models recommendation without actually considering the cases at hand. This over-reliance effect posed obvious concerns as there were still latent risks identified by the alerts. However, beyond the performance degradation, this situation raises concerns in terms of human oversight and accountability. This is an instance where introducing AI in human decision making might have an adverse effect and potentially reduce productivity.

Sample Use Case 3: Personalized Financial Services

AI-driven intelligence augmentation can significantly contribute to providing personalized financial services. By conducting a thorough analysis of an individual's financial history, spending habits, and future objectives, AI systems can provide tailored financial advice. This advice could range from suggesting suitable investment opportunities to providing effective budgeting and saving tips. AI chatbots can respond instantly to customer queries, enhancing customer service efficiency. They can also proactively alert customers about significant events, such as impending payments or changes in their account status. However, potential issues surrounding privacy and data security arise within this case study. Financial data is inherently sensitive, and any inadvertent disclosure could lead to serious repercussions. Consequently, financial institutions must implement robust data protection measures when employing AI.

Education

Sample Use Case 4: Personalized and Dynamic Learning Paths

The education industry is often underrepresented when discussing AI, but has a lot of potential for improvement and augmentation given current labor shortages. Leveraging generative AI can enable customization. By shifting focus from a one-size-fits-all learning approach to personalized learning pathways, the potent combination of watsonx.ai, Vela, and Hugging Face can drive transformative applications for AI in education, centered around augmenting human intelligence.

For example, a cutting-edge educational platform can be designed to collect student performance metrics (considering data collection protections for minors) based on questionnaires and surveys concerning individual learning styles, preferences, and growth plans. This platform can distill these complex datasets into tailored curriculum recommendations and resources, adapting to the unique needs of each student and making learning dynamic. As student requirements shift over time, so too can the educational model. The platform can evolve in tandem with the students, ensuring education remains relevant and engaging and, most importantly, freeing teachers up to do other important tasks.

Ethical considerations should be at the heart of this platform. Respect for students' privacy and data use are not afterthoughts, but integral components of the AI workflows facilitated by the watsonx.governance toolkit resulting in a student-centric, dynamic, and ethical learning experience. Teachers and students can benefit from a uniquely tailored education program supported by an AI system that understands and adapts to their needs. In this context, AI serves as a tool that enriches human potential.

Public Sector

Sample Use Case 5: Human Friendly Automation at German Employment Agency

The socio-economic challenges posed by AI-based automation served as the impetus for establishing the Human Friendly Automation (HFA) initiative in Europe. This cross-company endeavor aims to support employees and employers in

navigating the potential disruptions caused by AI-based automation, which can profoundly impact work dynamics and job sectors.

In collaboration with the German Employment Agency, the federal agency responsible for securing employment and labor market security in Germany, this initiative's research-based methods were put into practice in a groundbreaking project. The agency wanted to implement AI-based automation to enhance operational efficiency in their customer service center, employing AI for tasks such as automatic mail routing, text recognition, information extraction, and automated responses. By adopting the framework of the HFA initiative, the implementation process for this project underwent a fundamental change when compared to a typical IT implementation project. From the outset, interdisciplinary cooperation became essential with the project team, the client's IT and HR department, and change management demonstrating heightened attentiveness to employee needs.

A key outcome of our approach was that employees impacted by automation felt valued and included in the process. Our close collaboration with employees enabled us to identify their interests in pursuing new areas as technology assumed a greater role in their work. These valuable insights and our partnership with the HR team and CHRO facilitated the successful design and offering of new deployment opportunities and qualification programs.

We firmly believe that the successful scaling of intelligent automation technologies, both in the private and public sectors, hinges on preserving employees' wellbeing and ensuring clarity in their career paths, opportunities for up-and reskilling, as well as job transfers. This experience with the German Employment Agency exemplifies IBM's commitment to achieving this balance while harnessing the benefits of automation.

Healthcare

Sample Use Case 6: AI Assisted Patient Discharge and Bed Management

There are many areas spanning across the healthcare industry where AI can be infused. In this particular use case, AI can enable the prediction of patient discharge needs and efficient bed management in hospitals.

Machine learning algorithms can analyze extensive patient data, including electronic health records, vital signs, laboratory results, and clinical notes to discern patterns and trends related to discharge readiness. This comprehensive analysis includes factors such as the patient's diagnosis, treatment plan, comorbidities, length of stay, and overall progress. This approach allows healthcare professionals to prioritize patient care needs and expedite patient discharge, streamline operations and enhance overall workflow efficiencies, and proactively manage bed availability.

However, it is risky to allow AI-driven decisions to take precedence over those of healthcare professionals, especially regarding patient health. AI should be used to augment healthcare providers' decision making, supplying complementary information and insights to bolster quality of care. The inherent sensitivity of health data means that misuse and disclosure could result in significant consequences. As such, healthcare providers using AI should establish strong safeguards for data protection.

Sample Use Case 7: Early Detection of Cancer using AI

AI can support the analysis of medical imaging data, like mammograms for breast cancer or CT scans for lung cancer, to assist radiologists in the early detection of tumors. By analyzing large amounts of data and identifying patterns that may be difficult for humans to detect, AI can assist in the early detection of cancer. Sophisticated machine learning algorithms can enable AI to swiftly and accurately identify suspicious areas or lesions that require further investigation. By learning to recognize malignant tumor patterns, AI can provide a valuable preliminary assessments or second opinions for radiologists. This collaborative approach between AI and human experts supports a more accurate, efficient diagnosis, which can reduce false negatives and positives.

However, neither humans nor AI are universally correct in all medical diagnoses. Errors can lead to serious, life-threatening consequences. It can also be argued that AI-infused healthcare workflows could reduce the use of human intelligence, to the extent that decision making accountability is shifted from the human to the AI. AI could create bias in the doctors' original understanding and point of view, leading to different outcomes, for better or for worse.

Key Performance Indicators

Considering the various ethical concerns and implications of augmentation raised in these use cases, a framework of Key Performance Indicators (KPIs) may help measure the augmentation of human intelligence. Below, there are 10 identified measures that can be used as guiding metrics for evaluating an AI system's performance in relation to augmenting human intelligence.

KPI	Measurement	Pros	Cons
Learning capability amplification	Time taken to learn new tasks or adapt to changes in processes or environments before and after AI augmentation can be measured.	Could lead to an adaptable and agile workforce.	Learning and adaptability can be difficult to measure and can vary significantly among individuals.
Decision making improvement	Assessing the quality of decisions made by considering the outcomes. Surveys, feedback, and outcome analysis could be helpful tools.	Could lead to better strategic and operational outcomes.	Quality of decisions can be subjective and hard to measure accurately; impact of good decisions might not be immediately apparent.
Problem solving enhancement	Comparing the complexity and quantity of problems solved before and after AI augmentation, taking time into account.	Could lead to innovation and improvement in operations.	Complex to compare problem solving capabilities due to the unique nature of each problem; AI augmentation might not be beneficial for all types of problems and synergies.
Productivity increase	Track output over time before and after AI augmentation. Could be done through regular performance reviews and productivity tracking tools.	Straightforward to measure; increased productivity can lead to higher profitability.	Focus on productivity might lead to neglecting other important factors like job satisfaction or creativity; might not reflect the quality of work.
AI adoption rate (voluntary)	The percentage of employees effectively using the AI systems in their workflows.	Helps understand how effectively the AI is being integrated into daily operations.	Does not provide insights on the quality or impact of the AI usage.

KPI	Measurement	Pros	Cons
Task efficiency	Comparing task completion time and accuracy before and after AI implementation. This could involve direct observation, time-tracking tools, and accuracy-check mechanisms.	Easy to measure; immediate and tangible benefits can be observed.	Might not reflect long-term benefits; should avoid overly focusing on speed which could compromise quality.
Error reduction	Track the number of errors or quality issues before and after AI augmentation. This can be done via quality assurance reports and customer feedback. Tracking the source of error as human-driven vs AI-driven could also improve error reduction.	Clear measurement; reducing errors can lead to improved product/ service quality and customer satisfaction.	Some errors may be hard to quantify; reduction in errors may not directly translate to improved productivity or cost savings.
Cost savings	Comparing operational costs before and after AI augmentation. Look at specific line items that AI directly impacts such as labor costs, error correction costs, etc.	Easy to measure, direct impact on the bottom line.	May not account for intangible benefits or costs; upfront costs of implementing AI can be high.
Customer satisfaction	Monitor changes in customer satisfaction after AI augmentation through, for example, customer surveys or customer reviews.	Direct impact on business success; can improve brand reputation.	Can be influenced by factors outside of AI augmentation; customer satisfaction can be subjective.

Guidance

In order to put the principle of augmenting human intelligence into practice, we recommend the following best practices:

1. Use AI to augment human intelligence, rather than operating independently of, or replacing it.

2. In a human-AI interaction, notify individuals that they are interacting with an AI system, and not a human being.

3. Design human-AI interactions to include and balance human oversight across the AI lifecycle. Address biases and promote human accountability and agency over outcomes of an AI systems.

4. Develop policies and practices to foster inclusive and equitable access to AI technology, enabling a broad range of individuals to participate in the AI-driven economy.

5. Provide comprehensive employee training and reskilling programs to foster a diverse workforce that can adapt to the use of AI and share in the advantages of AI-driven innovations. Collaborate with HR to augment each employee's scope of work.

Conclusion

AI that augments human intelligence supplements the capabilities, experiences, and insights of human beings, balancing human oversight across the AI lifecycle. The use of AI to enhance human intelligence, rather than operate independently of or outright replace it, has the potential to transform how we address business challenges and make informed decisions. AI becomes a technological ally, activating human potential by empowering individuals to navigate a competitive business environment.

AI that augments human intelligence maintains human responsibility for decisions, even when supported by an AI system. Humans are therefore upskilled – not deskilled – by interacting with AI.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the
United States of America
November 2023