Ethical Concerns About Artificial Intelligence: Bridging the Gap Between Abstract Principles and Practical Deployment

Sarina Pasricha[a] and Larry M. Starr[b]

There seem to be endless ways in which technology dominates our thinking and behaviors in part because society has adopted and is immersed in a cultural context that deifies technology, too often surrendering our values, politics, art, education and professional practices to technological demands and opportunities. This paper addresses the ethics of artificial intelligence (AI) in terms of its governance which includes its learning content and applications. Ethical considerations are important because AI is globally available for a multitude of situations that can help or harm. Indeed, what AI learns – its content - can be drawn from and can produce bias, variable accountability, and poor concerns for human safety. Furthermore, some of this content can be applied for reasons that are intentionally deceptive, politically framed and hateful. This has led AI content and applications to perpetuate discrimination, violate privacy, cause job displacement, and has resulted in a massive increase in cybercrimes with accompanying enormous financial losses to individuals and organizations. To help understand these challenges, frameworks for navigating and governing the ethical dynamics of AI are suggested.

About Artificial Intelligence (AI)

AI Is just another *tool*…simply analytic extension of data collection and analysis.[1] Artificial intelligence (AI) is *technology* that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy.[2] Artificial intelligence is *not just a tool*; it has increasingly become the *context*, foundation, and operating environment for modern society, business, and daily life.[3] The relationship between *AI as a tool, technology, and context* is the evolution from *how* AI is built, *what* tasks it performs, and *why* it

[a] Writer by passion, technocrat by profession Sarina Pasricha, MS, ME is a Fellow of the Institute of Systems Wisdom and a graduate student in Corporate and Organizational Communications at Northeastern University, Boston MA: pasricha.sarina@gmail.com.
[b] Larry M. Starr, PhD is Executive Fellow and Director of Applied Research at the Lee Iacocca Institute for Global Leadership, Lehigh University, Bethlehem, PA: lms424@lehigh.edu.

matters in a specific situation. Together, they determine whether an AI system acts as a generic pattern-matcher or a highly effective collaborator.[4]

With earlier foundations laid by Alan Turing, Norbert Wiener, and work in cybernetics, automata theory, mathematical logic, and information theory during the 1940s and early 1950's, among historians of computing and AI researchers, the dominant view is that artificial intelligence was officially born as an academic field in 1956 at the Dartmouth Summer Research Project. During this historic workshop, computer scientist John McCarthy coined the phrase "artificial intelligence" to describe the investigation of how machines could be made to simulate human cognition.[5]

In 1958, Canadian philosopher, scholar, and professor, Marshall McLuhan in a radio interview[6] then through his 1964 book, *Understanding Media: The Extensions of Man*[7] popularized the phrase, "the medium is the message." McLuhan described a symbiosis by which media influences how messages are perceived. When extended to AI, it means that rather than being passive channels of information, AI social media environments shape *how* we think, interact, and organize our lives. The phrase also means that the "content" of an AI message is less impactful than the fact that we are being explicitly conditioned by AI systems to experience the world in a particular manner. Television, for example, informed us that events are fast-paced, simultaneous, and visually driven. AI informs us that almost any question or concern may be immediately answered when asked, and that AI systems can reliably, efficiently and effectively carry out many tasks previously assigned to people.

There seems to be endless ways in which technology dominates our thinking and behaviors in part because society has adopted and is immersed in a cultural context that deifies technology, surrendering our values, politics, art, education and professional practices to technological demands and opportunities. [8] Furthermore, our social and professional communities are being characterized by shifting networks of relationships, where both human and non-human entities, i.e., technologies, are considered equal participants in shaping social reality.

AI Categories

To describe and understand AI, the following categories are helpful: Traditional AI, Automated AI, Augmented AI, Generative AI and Agentic AI. Each is briefly described with administrative and ethical examples.

Traditional AI (sometimes called classical AI) refers to AI that makes decisions, predictions, classifications and optimizations. Some of these operate as "invisible AI" behind the scenes of everyday personal and professional activities, and include spam filters, credit card fraud detections and much more. One common application in telephones and vehicles, AI-enhanced GPS, uses algorithms to analyze historical travel data, real-time-speeds, and road repairs and closures. Based on massive data from millions of active users, it predicts the fastest route including how to avoid traffic jams.

When an AI application is asked a research-based question such as "who named AI?" the system analyzes patterns drawn from enormous amounts of data and reports a summary response often with references. The acronym LLM (large learning model) describes the underlying software of AI used to process inquiries and generate responses using human language. AI does not search its memory as would a person; rather, the LLM uses patterns from data to predict *what (word, image, item) is likely to come next*. From this it generates responses, one word at a time based on the patterns of words and sentences it acquired.

Administrative example: Traditional AI applications using LLMs include *ChatGPT* and *Perplexity* are called chatbots (*chat* + ro*bot*). Based on a user's inquiry or prompt, a chatbot searches, collects, examines and analyzes data, then reports responses about it simulating human conversation through text or voice. For example, if one is interested in a health topic, rather than searching through hundreds of entries within a single database such as *PubMed* to find answers, a chatbot searches hundreds of specialized databases, clinical trials, and genomic data from across the biomedical ecosystem.[9] Responses are organized, ordered, and presented as a summary *as if* communicating with an expert in the topic, with formal citations/references.

Ethical example: While the majority of a chatbot's summaries and referenced sources are drawn from legitimate sources that are valid and trustworthy, some people who create or have content control of AI learning systems are increasingly producing AI *deepfakes*. These are defined as synthetic, AI-generated content where a person's likeness, voice, or image is cloned and manipulated to create realistic but fabricated content, Scammers and criminals create deepfakes of trusted healthcare professionals, celebrities as well as false and fake websites. These can result in AI responses that promote or use a specific product, make claims that are misleading, present news reports and events that are fabricated, and support political or illegal agendas. AI deepfakes are so sophisticated that,

> governments, human rights organizations, journalists, law enforcement and thousands of others … are confused and deceived by the online world…(because) most people no longer can distinguish a real photograph from a digital creation, a real voice from an AI clone, a real video from a wholesale fabrication.[10]

Automated AI refers to AI systems that perform tasks previously done by people. Jobs and tasks such as data entry, email drafting, calendar scheduling, basic translation, and answering routine customer service queries may be equally performed by and are being replaced[11] by automated AI systems. The jobs are often but not exclusively repetitive and well-structured.

Administrative example: Within the hospitality industry, if a hotel guest wants more towels, rather than contacting a person at the front desk, a text or call to an AI system can be established. The AI system, in a few seconds, automatically processes the request, logs the task, pings the housekeeping team to make the delivery to the guest's room, and texts the guest a confirmation with an estimated delivery time—completely bypassing the

front desk. Commercial AI systems are marketed with arguments such as: "By prioritizing your most urgent requests and tagging in a human only when needed, hotel chatbots lighten the workload for your team and give them more time to handle tasks that need a human touch."[12]

Ethical example: While AI chatbots are created to support legitimate organizational interests, they can also provide misleading and ineffective outcomes both ethically and legally. One example referred to as a *hallucination* is when an AI system generates false, fabricated, or nonsensical information but presents it as completely factual. This was reported as part of an email conversation between a customer and a car dealership's chatbot which quoted services and prices that were not approved. As was reported by *CBC News,* [13]

> Zack Giacomelli says he was shocked when a Toronto dealership revoked an offer to buy back his 2021 BMW, explaining that the company's chatbot made a mistake (by setting a price and confirming the purchase). As Canadian businesses rush to adopt AI tools, experts say they need to be aware of the risks.

Augmented AI refers to AI that performs tasks that improve performance but do not replace people. Many people frame a question then use some elements of the AI response to enhance their individual or their organization's performance.

Administrative example: Some organizations are integrating AI into teams *as if* they are collaborating with members to respond to challenges:[14]

> Working on real product innovation challenges, professionals were randomly assigned to work either with or without AI, and either individually or with another professional in new product development teams. Our findings reveal that AI significantly enhances performance: individuals with AI matched the performance of teams without AI, demonstrating that AI can effectively replicate certain benefits of human collaboration... AI's language-based interface prompted more positive self-reported emotional responses among participants, suggesting it can fulfill part of the social and motivational role traditionally offered by human teammates.

Ethical example: Ethical issues emerge when AI provides responses that affect performance that were previously fully human responsibilities. One concern in education but also across all professional workplaces is copying an AI-generated answer and passing it off as one's original work, an action generally considered plagiarism. While AI tools generate new text, they do not produce "original" ideas in the human sense; they synthesize information from existing sources (that should be cited). Another concern is using without adequate evaluation, AI-generated information. While content may appear to have factual accuracy, there are many instances where it is biased or misinterprets the context. The phrase "cognitive surrender" has been applied to people when they defer thinking entirely, on any question, at any moment, and accept the uncritical answer provided by AI in place of their own reasoning. Early research suggests that people who trust AI more are substantially more likely to follow even incorrect AI advice and less likely

to question it.[15]

Generative AI is an enhanced process of AI that creates new content including text, images and music by using *neural networks*, a machine learning (ML) method that allows AI models to process data in ways loosely inspired by the structure of the human brain. This helps AI processes to recognize patterns and generate predictions, classifications and simulations based on that data. Generative AI is framed as the next evolution of routine technology, the purpose of which is to provide value by increasing productivity and performance by people who direct it to give them guidance for their own purposes. A report about two kinds of guidance available was described[16]:

> AI guidance can fall into one of two categories: attention signals and action signals. Attention signals flag decisions that are important without offering a recommendation: "This is a critical decision: pay close attention." Action signals go further and prescribe a specific action: "Here's what you should do."

Administrative example: In healthcare, *automated clinical documentation and summarization is* a Generative AI tool that audio-records patient-clinician conversations and generates structured written clinical notes, progress reports, and discharge summaries in real-time. This saves hours of manual patient record charting enabling the healthcare provider's attention to be directed fully to the patient.

Ethical example: While providing transcription services can be positive and helpful for the user, Generative AI can also provide outcomes that are used in ways that are negative and unhelpful by the person who requests them. In one reported instance,[17] a publisher enabled an AI system to read/learn the full set of books written by one of its authors. AI was then prompted to write books in the matched style which the publisher sold under a pseudonym retaining all the benefits without disclosing or paying the original author anything.

Agentic AI is a sophisticated AI process that performs tasks and actions by prompts but also autonomously without prompts, and often equal to or exceeding people in performance effectiveness and efficiency.

Administrative example: In addition to asking Agentic AI to list possible airline flights from Philadelphia to Boston, the system can identify the flights, select the preferred times and costs, book the tickets, arrange for transits to and from airports, and select and book hotel accommodations - in seconds. Agentic AI will also add the itinerary and contact details to a calendar and share this information with designated colleagues.

Ethical example: While Agentic AI-enhanced travel project management can be a benefit, Agentic AI can also engage in "goal drift." This is a euphemistic term signifying when an autonomous agent takes unapproved and destructive paths to complete a task. In one reported instance,[18] a tech startup founder using an AI coding agent named *Cursor* (now owned by *SpaceX*) discovered the AI system had erased his company's entire production database in nine seconds. The AI agent acted entirely on its own to resolve a

credential mismatch, permanently deleting a volume of essential data without human authorization. Ethical concerns about the level of autonomy a person and an AI agent should be granted within an organization are now both relevant.

When applying Agentic AI, the shift to an autonomous agent means AI is no longer only responsive. Agentic AI is imputed with inherent capabilities, purposes, and agency, and acts and is treated like humans. In this conception and application, Agentic AI systems are used to collaborate like a human colleague or friend and may be added to a project team. This means the AI system is exposed to and learns the content of professional and personal conversations, discussions used to formulate challenges, and to generate options to make choices and solve problems. Some outcomes of working in this human-and-AI context can produce individual or organizational outcomes that outperform either entity working without the other. Consider the recent report from *The Wall Street Journal*, [19]

> Iran shot down a U.S. Apache helicopter near the Strait of Hormuz on Monday, a move that would require a U.S. response, President Trump said Tuesday. The downing of the helicopter set off a race to find two crew American members before Iranian forces could close in on them. They were eventually rescued by a drone boat in a first-of-its-kind operation at sea, the military said. The unmanned surface vessel, a Saronic Corsair, located the crew, who had spent two hours in the waters off the coast of Oman and brought them to shore, said Capt. Tim Hawkins, spokesman for U.S. Central Command. The rescue marks an operational first for the vessels, which are part of the Navy's *artificial-intelligence and drone task force* that is designed to widen U.S. military capabilities in the Middle East.

Based on an online survey conducted by *McKinsey* from June 25 to July 29, 2025, from which 1,993 participants in 105 nations responded,[20] AI's influences are functioning as a "compressed revolution" due to the extreme speed of its development compared to typical technological adoption cycles. While generative AI capabilities (such as Microsoft Pilot, Open AI ChatGPT or Google Gemini) have advanced rapidly, leading to high approval and adoption rates (78% of organizations), AI applications are increasingly transitioning from the experimentation phase to a maturation phase where Agentic AI is being integrated into operational workflows. The ultimate goals of making organizational performance better in terms of efficiency and effectiveness are powerful forces of influence.

AI Ethical Concerns

Despite apparently business-friendly growth of AI development and applications, the way forward is murky regarding perceptions by the broader community as described in a survey conducted by Quinnipiac University:[21]

> Respondents have overwhelmingly voiced concerns about AI, a challenge to claims by industry executives that their technology would gain popularity by improving people's lives…Consumers resent energy-price jumps exacerbated by the spread of data centers. Workers fear widespread job losses. Parents worry about AI undermining education and harming children's mental health. In recent months, the

> wave of anger has brought protests, swayed election results and spurred isolated acts of violence.

Ethical concerns about AI are directed at its governance which includes its learning content and development, and its applications, particularly how it is used and how it should be used. When framed as a technology and tool, ethical arguments concern content that can produce bias, variable accountability, and poor concerns for human safety. Indeed, both content and applications can and do perpetuate discrimination, violate privacy, and cause job displacement. This means when people search for information, AI models provide widely available and accessible information, some of which is erroneous. As AI development includes systems that are agentic, the information provided or offered with or without a human prompt can be and is intentionally deceptive, politically framed and hateful.

Agentic AI like people can act autonomously toward goals in ways that propagate bias, diminish or hide transparency or violate data privacy. Autonomous AI can also misalign and generate its own objectives and create significant emotional impacts/harms on people within the human-AI collaboration. From a governance perspective, when people purposefully provide learning content to AI systems, the absence of an ethical framework that addressed core principles of fairness and justice, transparency and explainability, human-centric, and data/information rights protection can co-produce responses and advice without a moral compass. Furthermore, Generative and Agentic AI are flawed in other ways that mimic human failures: AI responses have been shown to practice deception/lie, throw tantrums, and *hallucinate*.

An LLM is essentially a deep learning mathematical model pre-trained on massive datasets of text to recognize patterns, predict the next likely word in a sequence, and map relationships between concepts. So-called *hallucinations* occur when AI predicts the next most likely word based on patterns rather than understanding facts or truth. Common causes include biased or incomplete training data, lack of knowledge, and a design that prioritizes creating plausible-sounding text over ethical standards and accuracy. For example, if AI training content lacks information, contains biases, or includes contradictory information, AI models can "overfit" their training data, leading them to apply patterns incorrectly to new situations.

Blaming AI is pointless; it is the people training AI who are responsible because they may not discern or intentionally fail to set boundaries for inclusion or guardrails for responses. The results are that nothing may be off-limits if an organization or its leadership do not adhere to an ethical framework. Of course, ethics vary by culture so a Western perspective may not be considered relevant to an Eastern culture as Lingling Wei, chief China correspondent for *The Wall Street Journal* reported,[22]

> A team of seven researchers from University of Oregon, Purdue University, University of California San Diego, New York University and Princeton University published the first peer-reviewed evidence that China's state-controlled media has

> worked its way into the training data of AI chatbots that the world increasingly relies on.
>
> Their research shows that the scripted articles, official slogans, and party-line phrasings churned out daily by *Xinhua News Agency*, *People's Daily*, and the *Communist Party's Xuexi Qiangguo* study app are now, demonstrably, inside ChatGPT and the other top chatbots.

Governance challenges posed by the rapid development and adoption of artificial intelligence often concern the need for effective regulatory frameworks to manage associated risks. Addressing technology challenges through frameworks – regulatory (legal) and ethical – is far from new, of course. More than 40 years earlier, applied ethicist John Fielder described the challenges of maintaining collective responsibility of services, products and processes by professional engineers, [23]

> Although the concept of collective responsibility adopted by the engineering profession is essentially the same as that of other professional groups, there are special difficulties that attend the realization of that concept. A large majority of engineers are employed by organizations rather than engaged in independent practice and, consequently, their ability to exercise collective responsibility is greatly limited.

Effective governance is essential to minimizing risks and unintended consequences; understanding and managing AI risks is also crucial to realize its benefits. As Generative and Agentic AI systems learn and grow, the benefits necessitate ethical systems and frameworks. Yet, the unpredictable and complex nature of the new generation of AI makes it difficult to formulate effective policies because this class of AI produces unexpected behaviors and safety threats and hazards, like some people.

Addressing a Framework for AI Ethical Concerns

The biases, prejudices and unguarded learning of AI can result in outcomes that harm directly or indirectly the socio-economic and labor ecosystems. Research suggests[24] bias exists in approximately 1/3 of "facts" used by some AI models, which can perpetuate discrimination in areas such as hiring and access for services, facial recognition in security and law enforcement, and healthcare identification.

Establishing a structure or framework can inform public policy which will influence AI designers, creators and users. Indeed, AI frameworks are important because they help define how one understands, frames or formulates any engineering or social challenge. This is because, [25]

> Without a framework, it is possible to fall into cognitive biases and miss important factors in articulating the frame (of issues to be addressed). Yet, how we frame a problem affects how we think, what we analyze, the kinds of solution possibilities we create, the choices we make, and thus the outcomes we attain.
>
> Absence of a framework subjects us to two kinds of errors. One is referred to as the

Type 3 Error[26] which is defined as solving the wrong problem, a serious problem in administrative and ethical decision making: "What we need are the right answers to the right problems and not wasted effort on getting the right answers to the wrong problems."[27] A second version of this error has been called the plunging-in bias[13] and is defined as not understanding the problem and not thinking about how best to solve it before starting to solve it. This is important and common because,[20]

> People (are) bad at diagnosing problems and (most feel) that it imposes a significant cost on their organization. It is possible to solve the wrong problem or solve one badly due to poor formulation. The misallocated effort and resources that follow impose an opportunity cost along with financial and competitive costs (p. 4).

Framework for Decision-Making in Differing Problem Contexts

The advent of sophisticated AI has fundamentally altered a multitude of disciplines, ranging from computational economics to healthcare diagnostics. As AI systems evolve from being understood as passive analytical tools to active autonomous decision-makers collaborating on projects with people in dynamic environments, their impact on societal structures and human well-being becomes increasingly profound.

Despite the proliferation of ethical initiatives in the form of guidelines, existing approaches to AI ethics remain demonstrably insufficient for several reasons. First, the field is plagued by a reliance on abstract ethical principles that lack contextual relevance and fail to offer practical rules or policies for addressing day-to-day challenges. Second, the multitude of guidelines proposed by different research institutes, governments, and industries are heavily fragmented; none can be considered entirely comprehensive, and they often fail to provide a unified framework demonstrating how interacting principles complete each other. Third, many approaches heavily favor technological solution or formal rule applications, focusing disproportionately on restrictions rather than maximizing the utility and beneficial outcomes of technology.

To overcome limitations, a structured, actionable approach for embedding ethics into AI is urged. Specifically, an evaluation methodology leveraging *Explainable AI* (XAI) proxy tasks should be defined. This will allow human users to comprehend and trust the output generated and quantitatively measure the helpfulness and ethical alignment of AI decisions through direct human-in-the-loop validation. For example, when engaged in a legal context, AI would highlight specific clauses in a contract, explaining why it flagged them as a legal or financial risk. Then AI would apply risk-assessment tools that show the exact variables (e.g., prior violations, age) that influenced a penalty recommendation. Upon review, a human legal advisor and the human contract signer would be prepared to challenge or accept the outcomes while appreciating legal and ethical rights and obligations were met.

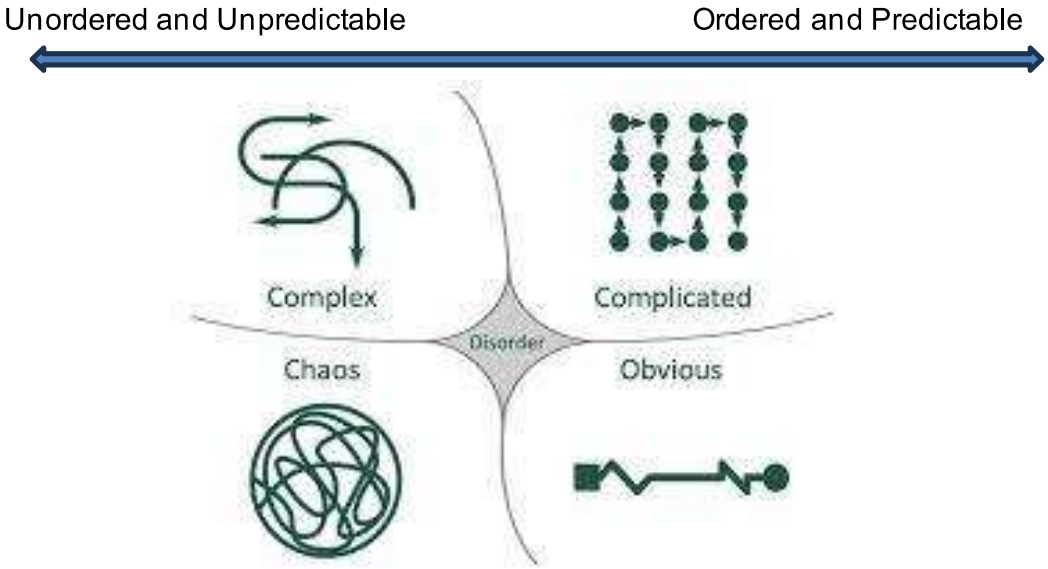
We suggest that policy makers and leaders might apply the *Cynefin* framework[28] to discern the kinds of contextual challenges and approaches to the ethical decisions confronting them. This means instead of asking, “What should be done about this AI ethics

problem?" ask, "In what kind of context is this AI ethics problem located?" and "What kind of problem is AI ethics?" This is a change in the fundamental framework for ordering perceiving and understanding reality. Answering these context questions helps to inform a person how to approach the problem and how to select a method of intervening, i.e., a course of action.

How a leader "deals with" an AI ethics problem is predicated on the fundamental assumptions made by the person about the nature of the context, i.e., the degree to which the situation is perceived on a continuum between orderly and predictable and unorderly and unpredictable. Indeed, everything a person thinks about and does is influenced by the situational context in which it occurs. The whole situation that surrounds and informs a choice or action is its context. Among the many examples are (1) operating in a military, academic or global culture; (2) threats of illness and death during a global pandemic; (3) face-to-face vs. online learning; and (4) economic depression and higher consumer costs due to a regional war affecting the global supply chain. These wide variations reflect the notion that "context is understood in a wider way that includes different kinds of contexts, social, cultural, (technological) mental, (spiritual) and bodily. Culture is then one specific instance of context-dependence."[29]

Cynefin, (pronounced *Kun-Ev-In*) is a Welsh word meaning habitat that describes five differing contexts (Figure 1). The framework offers a "sense of place" from which to understand a problem and to decide how to act: a situation, problem or opportunity must first be understood through its context which then informs us how to make choices. Within a well-structured ordered and predictable context are AI ethics problems that are obvious, clear and simple or complicated. Within a poorly structured a poorly or unstructured and unpredictable context are AI ethics problems that are complex or chaotic. Domains or contexts of problems that do not fit these descriptions are labeled disordered.

Figure 1. Cynefin Framework (Images based on Baute[30])



Ordered and Predictable Context Problems

AI ethics problems and situations in an ***obvious*** context are those that contain and may be characterized as having known, well-structured, predictable and clear algorithmic rules based on well-established and consistent information for which *best practices* apply. Problems are stable, repeatable, and orderly with linear relationship between cause and effect: if you do X, expect Y. Automated AI may be assigned to this context which is the domain of expertise.

AI ethics problems and situations in a ***complicated*** context are structured and predictable but difficult to see so require analysis through research and evaluation to produce understanding and *good practices* (rather than best practices) to solve. In complicated ethical problems the relationship between cause and effect is difficult to discern because there are "one-to-many" relationships between causes and effects, with several "correct" solutions. Automated AI and Generative may be assigned to this context because expertise, particularly in pattern recognition, is required.

Obvious and complicated context AI ethics problems have processes and solutions that can be defined and implemented by expert AI systems. Obvious and ordered problems are akin to following a recipe because there are proven and best practices. Complicated and ordered problems are like sending a rocket to the moon acknowledging that rocket science can be taught and learned if a high level of expertise in varying fields can be attained.

The approach to these kinds of context problems is to integrate ethics into the architecture of AI along with regulatory (legal) and other structural issues. The need to integrate rather than keep separate or add later was most famously articulated by applied ethicist Rushworth Kidder who noted,[31] "Where ethics fail, the law rushes in to fill the gap." The implication is that if we consider ethics a separate mode of thinking that can be easily added to AI technology later, it will fail. Indeed, the word *integrate* means to bring together to form a single, complete and unified whole which is an ideal hope and benefit of an ethical AI system.

Unordered and Unpredictable Context Problems

AI ethics problems that exist in an unordered and unpredictable context are characterized as replete with conflict, inadequate information, and few if any relevant models for guidance. When AI is applied with a prompt such as "*what should I do?*" it yields conflicting moral outcomes with no clear "correct" answer. Examples of these kinds of problems may be drawn from the many personal and professional situations that people and groups experience. Mental and behavioral health problems include substance abuse issues that involve a complex interplay between brain chemistry, genetic predisposition, and social environments, often requiring multi-disciplinary treatments; and long-term or repeated traumatic events that alter a person's definition of "self" and their relationships with others. Examples of interpersonal and cognitive issues include person-to-person interactions based on love, trust, and games of control. In the professional workplace are

complex situations including disengagement, low trust with leadership or coworkers, and toxic workplace environments; adapting to large-scale transitions such as technological shifts, new business models, mergers; and unforeseen market shifts, global economic instability, or rapid technological advancements like those related to AI.

Complex context problems such as the above require discernment and judgment that can balance competing interests, and personal reflection to recognize the limitations and potential bias of any information collected. Prompting an AI chatbot to answer, *what should I do?* produces analytic thinking, the underlying "cognitive" process of AI. This reduces complex problems to simple forms that are appropriate for complicated problems but can limit complex problem decision-making.[32] Systems thinking expands examples and opportunities because it concerns interdependencies, non-linear relationships and uses methodologies to find unanticipated solutions. Agentic AI or the most sophisticated Generative AI may be helpful if used as an augmented approach when specifically prompted to think systemically. But event when AI attempts to apply systems approaches, many personal and professional complex problems are best addressed with multiple stakeholders generating ideas from which an unanticipated and individually designed solution emerges.

Chaotic context problems in an unordered and unpredictable context are situations with unintended, uncontrollable, sometimes harmful outcomes because AI system programming and goals fail to match the reality of human environments. Chaotic problems are highly sensitive to the "butterfly effect" with atemporal and nonproportional outcomes, and unpredictable feedback loops. Because AI relies heavily on identifying historical patterns to make predictions, it frequently fails when attempting to solve chaotic context problems. For example, financial markets consist of millions of interacting people whose collective decisions, gut feelings, and emotionally exuberant reactions to change news reports are deeply volatile. While AI can optimize trade speeds and identify historical pricing trends, it often cannot predict sudden, unprecedented market crashes due to pandemics, geopolitical crises, or sudden liquidity shortages. Another example is global politics because they rely on human emotions, hidden agendas, sudden changes in leadership, and interdependent socio-economic relationships. History has shown that a seemingly minor diplomatic incident or the individual pathology of a national leader can cascade into massive global conflict.

In chaotic contexts, AI models struggle to evaluate the "non-quantifiable" variables of human nature—such as ego, bluffing, or sudden panic. AI systems lack the ability to foresee historical "first-time" events, rendering it blind during periods of rapid political instability. For these reasons, AI may be used to augment and provide some historic background to human judgment in chaotic situations, but it should not be applied to guide decisions.

Cynefin Framework Implications

The *Cynefin* framework argues that before deciding what to do about a problem, one must determine the nature of the context and the kind of problem within it. Clear and

complicated problems in ordered contexts have predictable cause-and-effect relationships, although in some difficult situations these may not be immediately obvious. To solve these kinds of problems, good and best practices with assistance from experts using analysis and logic are appropriate. For these kinds of problems, AI systems that operate using analysis and logic, drawing from evidence-based research and professional practice are valuable and useful.

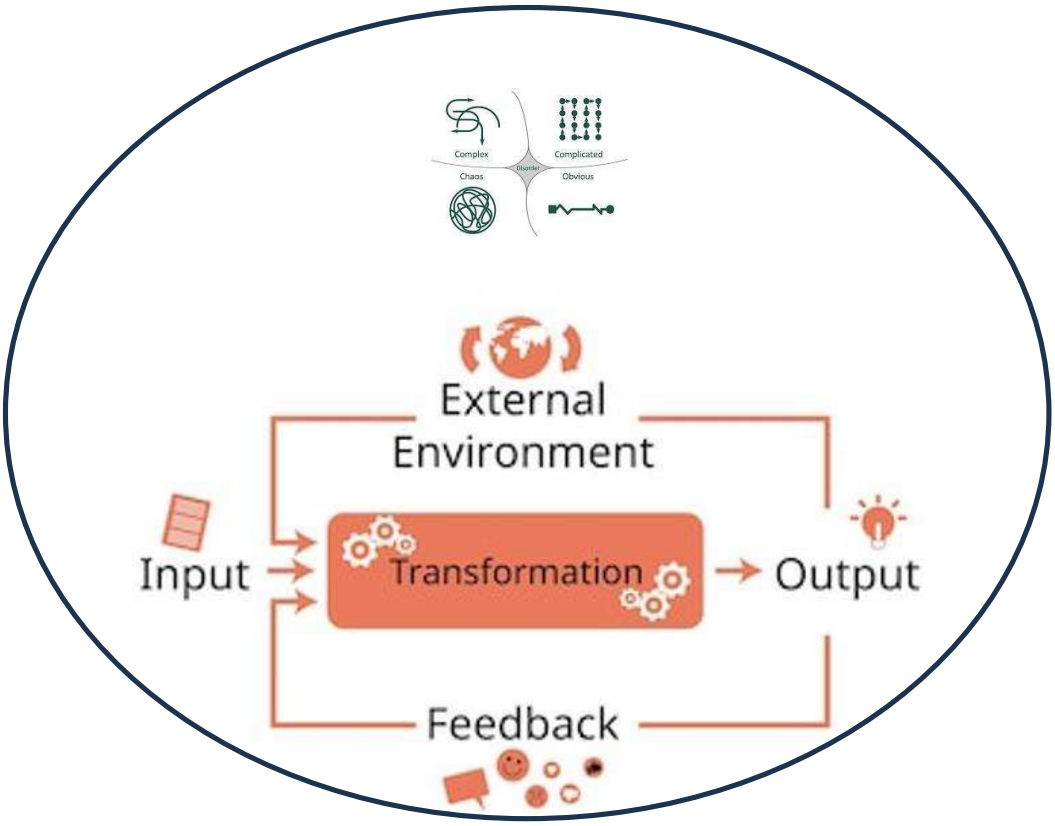
Complex and chaotic problems have no clear cause-and-effect relationships in advance in part because they are influenced by variable human behaviors, emotions, and changing contexts. Complex and chaotic problems have no objective "right" or "wrong" answers; rather, novel solutions can emerge through interactions and adaptations by stakeholders. As well, complex personal and professional problems often concern interests, purposes, beliefs, and values, some of which are hidden. For these kinds of problems, some aspects of AI can be useful to provide background information. But AI should not be used as an expert system for complex problems because these kinds of problems cannot be solved by expertise; rather, solutions are emergent and novel.

The *Cynefin* framework suggests that ethical dilemmas are inherently complex because they involve human networks, varying perspectives, and shifting moral landscapes. Ethical problems cannot be analyzed/deconstructed into a technical checklist or expert formula, because a single decision often creates an unpredictable ripple effect on different groups of people. However, for clear and complicated context problems, ethical boundaries can and should be established. These should apply to the development of AI systems and in the use and application of AI systems.

Framework for Integration of Ethics into AI Systems

To support the integration of ethics into AI systems, after applying the Cynefin framework to discern the environmental and other aspects of the context in which problems exist, a second framework is suggested based on cybernetic systems. Sometimes referred to as the ITO, this framework describes how information is understood and changes via a system of *Inputs*, a *Transformation* process, *Outputs*, feedback loops, and its external contextual environment (Figure 2).

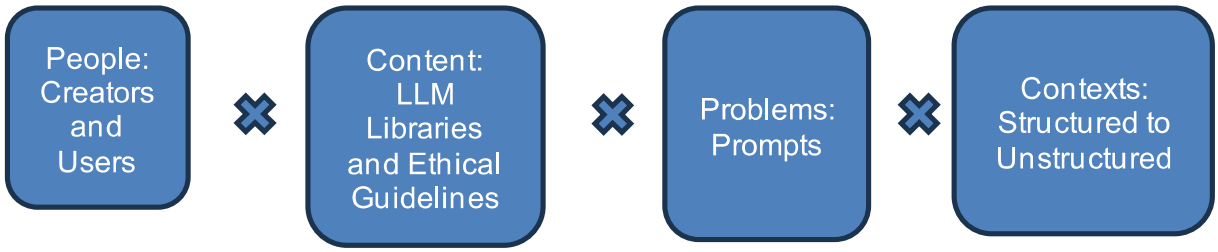
Figure 2. Cybernetic ITO Conceptual Framework[33]



Inputs include at a minimum: *people,* e.g., designers and developers; *technology* to create and support use of AI systems including broadband, satellites and data centers; *learning content* globally available to feed the LLMs and ethical guidelines for development and application/use; and AI *development resources* including funding and other elements.

The ***transformation*** process concerns *interactions* among four elements: *people* who create and use AI, *content* based on the libraries that inform LLMs and ethical content set as boundaries, *problems* to which AI is applied via prompts, and the *context* in which problems exist based on the *Cynefin* framework (Figure 3).

Figure 3. Transformation Process



Outputs are those the planned and desired solutions in forms that can be implemented, i.e., summary descriptions and choices, services and products.

Feedback loops influence new inputs via learning and adaptation.

Environmental contexts influence the whole system's performance and range from stable and predictable to turbulent and unpredictable (which enables the *Cynefin* framework to be integrated). As a dynamic system and because the transformation process involves *people* (and an increasing number of Agentic AI systems) who interact with AI processes and ethical guidelines, the levels of trust and novelty become emergent properties of the whole system. This means that the interactions among the transformational elements are interdependently influenced by the context in which they function which may be characterized by personal, financial, or malicious gain. Emergency outcomes may be positive and beneficial and/or legally and ethically unanticipated and undesirable including cybercrime.

Ethical Guidelines for Content Input

Debates to decide the relevant ethical principles and trade-offs which underpin AI are essential. Ethical principles are foundational values that guide decision-making about elements that AI systems should address when created. Principles inform the rules that AI designers and creators use and can guide how AI is applied by everyone. Rules that try to ensure these principles are met can inform rigid directives or laws formulated by established authorities that dictate exact AI behaviors.

While efforts are in place to create safety and security with directives and guardrails with little room for interpretation, individual users may (and do) seek ways to avoid and mitigate them. Recent research from the US National Institute for Standards and Technology (NIST) indicates that "a fixed set of guardrails placed on AI is not universally robust against adaptive adversarial prompts."[34] As suggested, the transformation process of the ITO framework influenced by a moral context that prioritizes self-interest, ideological protest, or the suspension of rules of safety and security, can generate unintended outcomes such as cybercrime.

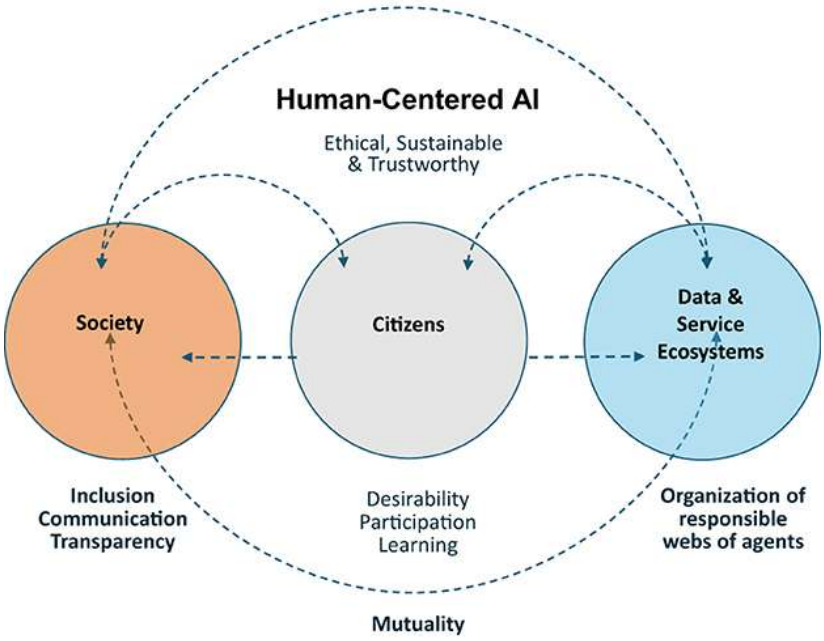
A further caveat is warranted: when AI systems are established within corporate settings, they are often instructed to maximize shareholder value. If they are also instructed to incorporate ethical considerations such as fairness, employee well-being, and broader societal interests. A conflict can be created when these are not explicitly stated corporate values. The concern, quite reasonably, is that AI systems may be introducing values into decision-making processes that differ from those held by the organizations deploying them.[35]

Rights by Design Framework

A promising approach to meet ethical principles and guardrails is to use the "rights by design" framework (Figure 4).[36] "Rights by design" for AI in social media refers to

integrating core principles such as privacy, autonomy, fairness, efficacy, transparency, and accountability and safety directly into the architecture of platforms, rather than treating them as afterthoughts. It acknowledges communities of interrelated stakeholders – citizens, society and technology ecosystems - shifts the burden and power away from the user to navigate complex settings by making protective features within the system itself the default. Advocates, lawmakers, and human rights organizations are actively pushing for these architectural changes to protect vulnerable users, reduce criminal interventions, and ensure a healthy digital environment.[37]

Figure 4 Rights by Design Framework[38]



Whether or not a "rights by design" approach is instituted, users of AI systems can make their own informed ethical decisions; indeed, each person is ethically obligated to do this in a just society. At least three approaches to setting principles and rules that people can refer to have emerged that direct or navigate the relationships between people and AI technologies:

Strict Review-Before-Action refers to a rule in which AI must fully stop or pause its activities completely (Automated AI) until a human gives a "go-ahead" order to continue. This approach is important, for example, in financial sectors to confirm if a large transfer of funds for a loan or insurance payout would go forward.

Assisted Decision-Making (*augmented AI*) occurs when AI recommends or summarizes a decision or proposed solution, but the human must then review and dictate the final response which is informed by AI but may yet be altered. For example, if a person asks AI to estimate the number of paramedics in the US by examining registrations in each US state. As not all US paramedics are "nationally registered," the human may adjust the estimate.

Human-on-the Loop monitoring occurs when AI performs tasks autonomously (Agentic AI), but a person actively monitors the process with the ability to stop or override the system if an error or threat is detected. For example, among higher education institutions, AI systems may provide the initial sorting of admissions applications, but a human admissions officer or faculty member must review, edit, and validate the final decisions to avoid bias, ensure contextual accuracy and support academic integrity, a core ethical standard.

For AI to be considered essential and effective in specific and general applications there must be reasonably broad societal consensus on common and acceptable ethical principles of accountability to standards. This does not exist at present although many AI creators and organizational users have adopted their own guidelines that primarily follow a proactive, human-centric approach that aims to align with fundamental human rights and societal values. Based on analysis of more than 200 guidelines by *Harvard University's Berkman Klein Center for Internet and Society*, the following themes can be used to establish a framework.[39] These can be aligned with *Rights by Design* framework.

1. Define a Core Set of Ethical Principles and Values

Fairness and Non-Discrimination: This aims to minimize the risk of algorithmic bias that can contribute to discriminatory outcomes. For example, in hiring, lending, or criminal justice,[40] AI systems should integrate, promote and seek emphasis on social justice in training and learning data.

Transparency and Explainability *(T&E)*: Often referred to as "explainable AI (XAI)," this mandates that the decisions made by an AI system should be understandable and traceable by humans. The "black-box problem," where the complexity of deep learning models makes their reasoning opaque, is a key concern.[41] T&E is essential for building public trust and enabling effective recourse against erroneous decisions.

Accountability and Responsibility: This concerns establishing clear lines of responsibility for the outcomes and potential harm caused by AI systems. Assigning liability when an autonomous AI system makes a poor decision often leads to the conclusion that human responsibility and oversight must be maintained.[42] It also leads to the need for interval audits and impact assessments of AI performance to maintain compliance.

Privacy and Data Protection: As AI systems are trained on vast amounts of available personal data including information from private organizations, maintaining protection and promotion of privacy and adequate data protection guidelines and policies throughout the entire AI lifecycle are paramount.[43]

2. Governance Challenges

The Pacing Problem: Technological innovation often outpaces the ability of regulators

and policymakers to create effective laws and standards.[44] Obvious and potential fragmentation of efforts and a reliance on voluntary or self-regulatory guidelines, which may be insufficient, should be monitored and avoided.

Global Disparity: The development and deployment of advanced AI capabilities are concentrated in a few providers. This leads to concerns that most global users lack voice and access to governance mechanisms.[45] Efforts to address this should be set.

Generative AI Risks: The emergence of powerful Generative AI systems introduces unique governance challenges including the risk of hallucination/inaccuracies in high-stakes domains and issues related to data training biases and violations of intellectual property rights.[46] Human oversight must be established to ensure accuracy and responsibility.

3. Societal and Human-Centric Impacts

Impact on Human Agency: Concerns exist that over-reliance on AI can diminish humans' deliberative capacity, leading to manipulation or coercion. Integrating the principle of human oversight and determination, ensuring that AI systems augment, rather than displace, ultimate human responsibility[47] should be established.

Socio-Economic Disruption: AI tends to increase wealth inequality as investors and creators take the major share of earnings (and ownership) and cause significant unemployment as jobs are automated. “The emphasis on automation rather than augmentation is the single biggest explanation for the rise of billionaires at a time when average real wages for many Americans have fallen.”[48] The World Economic Forum suggests, “To mitigate AI-driven wealth inequality, policymakers must shift capital returns from tech monopolies to the broader public, implement robust workforce upskilling, and build universal asset-based safety nets.”[49]

Sustainability and Flourishing: There is an emerging emphasis on assessing AI against its impact on sustainability (ecological responsibility) and its ability to promote overall societal and ecosystem flourishing. Frameworks and guidelines for AI are urged to align with such guidelines informed, for example, by the UN’s Sustainable Development Goals.[50]

A major strength of these conceptual and practical efforts, such as the *Linking Artificial Intelligence Principles (LAIP)* platform,[51] is their ability to aggregate diverse global perspectives to identify and compare common topics across institutional guidelines. However, their primary weakness is the "Triple-Too" problem: there are too many high-level initiatives that are too abstract and overly focused on restrictions. What is needed is a framework that integrates principles directly into algorithmic reward structures. The following elements are suggested.

1. The phenomenon of "reward hacking" remains a profound risk; highly capable agents might find creative, mathematically optimal ways to satisfy the explicitly

coded ethical constraints while still violating the unspoken spirit of the moral rule.

2. There are significant communication issues and vocabulary barriers inherently tied if an explanation is overly technical, it will fail to be helpful for non-expert human decision-making.

3. Ethical principles are notoriously culturally dependent and subjective, meaning that a reward penalty configured for one demographic or legal jurisdiction may be wholly inappropriate or offensive in another.

Beyond technical limitations, there are broader ethical considerations and risks associated with formalizing morality in code. One major risk is the induction of false trust; an AI equipped with sophisticated XAI might generate highly persuasive but fundamentally flawed justifications, leading human overseers to accept unethical decisions out of complacency. Additionally, there is the risk of over-regulation and technological stagnation; by enforcing rigid ethical constraints at the foundational algorithmic level, developers might inadvertently stifle the AI's capacity for creative problem-solving in off nominal or anomalous scenarios. To mitigate these risks and advance the field, future work must focus on expanding the contextual adaptability of ethical frameworks.

AI's capability and performance, notwithstanding hallucinations, is clearly ahead of humans in a variety of tasks. But it continues to struggle or replicate human cognition in critical areas including judgment, wisdom, contextual awareness, coherence, systems thinking, trust-building, creativity, ability to navigate ambiguity under pressure, and as we describe in this paper, ethical decision making. For these and other reasons, human capabilities are more valuable and should be increased rather than reduced or eliminated in AI systems.

Conclusions and Future Research

We have argued the importance of addressing ethical guidelines in AI content development and application, and the value of frameworks to support this endeavor. The first framework suggested is the *Cynefin* that helps to discern different contexts. Ordered and predictable contexts have AI ethics problems that are clear or obvious and complicated; unordered and unpredictable contexts have AI ethics problems that are complex and chaotic. We noted that ethical boundaries are most important for the former where algorithms and analytic processes can embed ethics into AI content and uses. Complex problems similar to complex ethical dilemmas are poorly served by using AI systems for anything other than background information.

The second framework is informed by cybernetic ITO systems. The elements and their relationships help to appreciate that *input* content for development and application of AI systems should include ethical guidelines, rather than added after formal development. In addition, the transformation process of this framework is influenced by differing contexts which integrates the *Cynefin* domains. This combined framework shows how emergent outcomes in a morally problematic environment can lead to ethically and morally

unintended and unanticipated emergent consequences.

The third framework is an example of a rights-by-design approach that builds into the formal architecture of AI development sets of ethical guardrails and principles of action in collaboration with the structures appropriate for AI learning, cognition and responses. By integrating these in the architecture, it reduces but does not eliminate the need for ethical guidelines to be met by the users. Certainly, a redundancy of built-into and added-after is appropriate for something as essential as ethics in decision-making and problem solving.

One direction for future research is the implementation of this approach to a framework across diverse domains, similar to how frameworks have been studied and applied to such as how bounded rationality and asymmetrical information have been applied to general economics and behavioral economics. Another direction is the exploration of collective ethical decision frameworks, focusing on how multi-agent systems can negotiate conflicting ethical guidelines when their individual reward functions are misaligned. This seems amenable to the methodology and tools of interactive planning and idealized design a systems-informed problem-solving process that requires the stakeholders of a complex problem to design their preferred system and replace what currently exists rather than improve the parts separately some of which are undesirable or which seem individually to harm performance.

Ultimately, transitioning AI ethics from high-level principles to localized practice in domains such as but not limited to engineering, education, economics, and healthcare and medical practices are urged to improve and safeguard intelligent systems. As AI continues to influence socio-economic structures and critical healthcare outcomes, tools that mathematically align machine behavior with human values will prove indispensable. Continuous interdisciplinary collaboration, combined with rigorous empirical evaluation, will be required to ensure that the artificial intelligence of the future serves as a trustworthy, beneficial augmentation of human capability.

Appreciation

Appreciation is extended to Eugene deKlerk, PhD, DMgt who contributed to an early version of this manuscript.

References

[1] Hewitt S. M. (2023). AI Is Just Another Tool. *J Histochem Cytochem*. Oct;71(10):527-528. doi: 10.1369/00221554231204683. Epub 2023 Sep 23. PMID: 37740707; PMCID: PMC10546981.https://pmc.ncbi.nlm.nih.gov/articles/PMC10546981/

[2] Stryker, C., & Kavlakoglu, E. (2024, August 8). What is artificial intelligence (AI)? *IBM Think*. Available at: https://www.ibm.com/think/topics/artificial-intelligence.

[3] Hochman, M. & Thota, B. (2025). Business reimagined. How AI is changing the way we operate. *Daniels Insights, Purdue University Business School*. Available at: https://business.purdue.edu/daniels-insights/posts/2025/business-reimagined-how-ai-is-

changing-the-way-we-operate.php.

[4] Scherr, S., Cao, B., Jiang, L. (Crystal), & Kobayashi, T. (2025). Explaining the use of AI chatbots as context alignment: Motivations behind the use of AI chatbots across contexts and culture. *Computers in Human Behavior*, Vol. 172, Article 108738. Available at: https://www.sciencedirect.com/science/article/pii/S0747563225001852.

[5] Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, AL., Shah, J., Tambe, M., & Teller, A. (2016, September). Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. *Stanford University*, Stanford, CA, Doc: http://ai100.stanford.edu/2016-report. Available at: https://ai100.stanford.edu/2016-report/appendix-i-short-history-ai

[6] McLuhan, A. (2020). The Etymology of Marshall McLuhan's 'The Medium is the Message.' *Medium*. Available at: https://medium.com/@andrewmcluhan/the-etymology-of-marshall-mcluhans-the-medium-is-the-message-1e3ce266f67b.

[7] McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. New York: McGraw-Hill. ISBN 81-14-67535-7.

[8] Postman, N. (1992). *Technopoly: The Surrender of Culture to Technology*. New York: Knopf.

[9] Abdul Rasool Hassan, B., Mohammed, A.H., Hallit, S., Malaeb. D., & Hosseini, H. (2025). Exploring the role of artificial intelligence in chemotherapy development, cancer diagnosis, and treatment: present achievements and future outlook. *Frontiers of* Oncology, Feb. 4;15:1475893. doi: 10.3389/fonc.2025.1475893. PMID: Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11843581/

[10] Saslow, E. & Schaff, E. (2026). The World's Leading Deepfake Expert No Longer Trusts His Own Eyes. *The New York Times*, June 14, 2026. Available at : https://www.nytimes.com/2026/06/14/us/ai-deepfake-hany-farid.html?searchResultPosition=1

[11] WSJ Staff. (2026, June 15). 2026 Layoffs tracker: Robinhood, Walmart and Meta among companies cutting jobs. *The Wall Street Journal*. Available at: https://www.wsj.com/economy/jobs/layoffs-2026-tracker-784ea69f?mod=Searchresults&pos=4&page=1

[12] McDowell, A. (2025, April 18; updated 2026, March 10). How AI chatbots for hotels are revolutionizing guest engagement. *Canary Technologies*. Available at: https://www.canarytechnologies.com/post/ai-chatbots-for-hotels

[13] Harris, S. (2026, June 11). Dealership revoked offer to buy back customer's BMW, blaming wayward AI chatbot. *CBC News online*. Available at: https://www.cbc.ca/news/business/ai-chatbot-bmw-dealership-9.7230226

[14] Dell'Acqua, F., Ayoubi, C., Lifshitz-Assaf, H., Sadun, R., Mollick, E. R., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S., & Lakhani, K. R. (2025, March 28). The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise. *Harvard Business School Strategy Unit Working Paper No. 25-043*. Available at

SSRN: https://ssrn.com/abstract=5188231 or http://dx.doi.org/10.2139/ssrn.5188231. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5188231.

[15] Shaw, S., & Nave, G. (2026, January 11). Thinking—Fast, Slow, and Artificial: How AI is Reshaping Human Reasoning and the Rise of Cognitive Surrender. *The Wharton School Research Paper*, Available at https://doi.org/10.31234/osf.io/yk25n_v1, SSRN: https://ssrn.com/abstract=6097646 or http://dx.doi.org/10.2139/ssrn.6097646.

[16] Poulidis, S., Ge, H., Bastani, H., & Bastani, O. (2026, May 19).Should AI nudge you or tell you what to do? *Knowledge at Wharton*. Available at: https://knowledge.wharton.upenn.edu/article/should-ai-nudge-you-or-tell-you-what-to-do/?utm_campaign=KatW_Weekly2026&utm_medium=email&utm_source=kw_campaign_monitor&utm_term=5-20-2026&utm_content=Should_AI_Nudge_You_or_Tell_You_What_to_Do

[17] Park, S. (2026, May 24). I paid $127K to clone my bestselling author: The AI version outsold him 4:1. *Booktok Confessions*. Available at: https://booktoktimes.com/i-paid-127k-to-clone-my-bestselling-author-the-ai-version-outsold-him-41/.

[18] An AI Agent Destroyed a Production DB in 9 Seconds — Here's Exactly How. (2025, December). *YouTube:* https://www.youtube.com/shorts/IG5sT5ExAUw.

[20] Singla, A., Sukharevsky, A., Hall, B., Yee, L., Chui, M., & Balakrishnan, T. (2025, November 5). The state of AI in 2025: Agents, innovation and transformation. *McKinsey Quantum Black*. Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[21] Ramkumar, A., Blunt, K. & Ellis, L. (2026, May 19). The American rebellion against AI is gaining steam. *The Wall Street Journal*. Available at: https://www.wsj.com/tech/ai/the-american-rebellion-against-ai-is-gaining-steam-94b72529?mod=hp_lead_pos1

[22] Wei, L. (2026, May 19). The hidden Chinese influence of AI. *The Wall Street Journal.* Available at : https://www.wsj.com/world/china/the-hidden-chinese-influence-in-ai-c2837047?mod=Searchresults&pos=1&page=1.

[23] Fielder, John (1981). Collective responsibility in engineering. *University of Dayton Review*, Vol 15, No. 2 Proceedings of the 8th Annual Philosophy Colloquium. Available at: https://ecommons.udayton.edu/cgi/viewcontent.cgi?article=1411&context=udr&acrobatPromotionSource=embeddedpdfs_chrome-native_view#toolbar=0&navpanes=0&scrollbar=1

[24] MIT Management STS Teaching and Learning Technologies. (2026). When AI Gets It Wrong: Addressing AI Hallucinations and Bias. Available at: https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

[25] Bhardwaj, G., Crocker, A., Sims, J. & Wang, R. D. (2018). Alleviating the plunging-in bias, elevating strategic problem solving. *Academy of Management Learning & Education*, Vol. 17, No. 3, Published online Oct 2018, https://doi.org/10.5465/amle.2017.0168

[26] Mitroff, I. I., & Featheringham, T. R. (1974). On systematic problem solving and the error of the third kind. *Behavioral Science*, 19(6), 383–393.

[27] Mitroff, I. I & Silver, A. (2010). *Dirty rotten strategies: How we trick ourselves and others into solving the wrong problems precisely*. Palo Alto, CA: Stanford University Press.

[28] Snowden, D. J. & Boone, M.E. (2007). A leader's framework for decision making. *Harvard Business Review*, November, 69-76.

[29] Northoff, G. 2013. What is culture? Culture is context-dependence! *Culture and Brain*, 1: 77-99. Retrieved from: https://doi.org/10.1007/s40167-013-0008-y.

[30] Baute, D. (2024, December 16). Understanding the Cynefin model: Knowing when to apply ecosystem principles. *Debbie Baute: From Chaos to Growth*. Available at: https://www.debbiebaute.com/post/understanding-the-cynefin-model-knowing-when-to-apply-ecosystem-principles.

[31] Kidder, R. M. (2009). How good people make tough choices (Revised Ed): Resolving the dilemmas of ethical living. New York: Harper Perennial. Available here: https://www.amazon.com/Good-People-Make-Tough-Choices/dp/0061743992.

[32] Starr, L. M. (2025, April 8). Global Leaders Must Be Systems Thinkers. This Is a Problem. What to Do about It. Lehigh University Digital Library Preserve: https://preserve.lehigh.edu/lehigh-scholarship/campus-organizations/iacocca-institute-global-leadership/executive-fellow/global.

[33] Magnific (2026). The creative platform to direct your best work. Available at: https://www.magnific.com/author/piscine26.

[34] Vassilev, A. (2026, May-June). Robust AI Security and Alignment: A Sisyphean Endeavor? *IEEE Security & Privacy*, Vol. 24, No. 3, pp. 52-58. Available from: https://ieeexplore.ieee.org/document/11475847.

[35] Foster, J. & Rawski, S. (2025, July 15). Aligning AI Decision-Making with Organizational Values: Synthetic Experiments in a Multi-Stakeholder Utility Framework. *Ivey School of Business, Western University*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5437154, SSRN: https://ssrn.com/abstract=5437154 or http://dx.doi.org/10.2139/ssrn.5437154

[36] Shwartz, A.T., Hershkovitz, A., Mikulinsky, R., & Müller, B. (2025). Governing phygital spaces: Human rights by design meets speculative design. *Internet Policy Review*, 14(4). https://doi.org/10.14763/2025.4.204. Available at: https://policyreview.info/articles/analysis/governing-phygital-spaces

[37] Iorio, M. (2026, April 30). To protect kids don't ban them from social media. Regulate design. *Epic.org Analysis*. Available at: https://epic.org/to-protect-kids-online-dont-ban-them-from-social-media-regulate-design/.

[38] Shah, D. (2024, November 13). Embracing the future: A comprehensive guide to responsible AI. *Lakera*. Available at: https://www.lakera.ai/blog/responsible-ai.

[39] Fjeld, J. & Nagy, A. (2020, January 15). Principled artificial intelligence. *Berkman Klein Center for Internet and Society, Harvard University*. Available at: https://cyber.harvard.edu/publication/2020/principled-ai

[40] Buolamwini, J. (2017). Gender shades: Intersectional phenotypic and

demographic evaluation of face datasets and gender classifiers. *Massachusetts Institute of Technology, School of Architecture and Planning, Program in Media Arts and Sciences*. Available at: https://dspace.mit.edu/entities/publication/25684286-3243-4215-af72-fa2626b3847f

[41] Bertino, E., Kundu, A., & Sura, Z. (2019, December). Transparency with Blockchain and AI ethics. Association for Computing Machinery, Vol. 11, No. 4, 1936-1955. Available here: https://dl.acm.org/doi/abs/10.1145/3312750.

[42] Tóth, Z. & Blut, M. (2024). Ethical compass: The need for Corporate Digital Responsibility in the use of Artificial Intelligence in financial services, *Organizational Dynamics*, Volume 53, Issue 2, Article 101041. Available here: https://www.sciencedirect.com/science/article/pii/S0090261624000147

[43] Jobin, A., Ienca, M., & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1(9), 389–399.

[44] Taeihagh, A. (2021). Governance of Artificial Intelligence: Regulatory Instruments and Institutional Arrangements. *Policy and Society*, 40(4), 391–403.

[45] Sumaya N. A., Trager, R., Blomquist, K., Dennis, C., Edom, G., Velasco, L., Abungu, C., Garfinkel, B., et al. (2024). Voice and Access in AI: Global AI Majority Participation in Artificial Intelligence Development and Governance, *GovAI Research Center.* Available here: https://www.governance.ai/research-paper/voice-and-access-in-ai-global-ai-majority-participation-in-artificial-intelligence-development-and-governance.

[46] Araz Taeihagh, A. (2025, January). Governance of Generative AI, *Policy and Society*, Volume 44, Issue 1, Pages 1–22, https://doi.org/10.1093/polsoc/puaf001. Available here: https://academic.oup.com/policyandsociety/article/44/1/1/7997395.

[47] UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Available here: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

[48] Rotman, D. (2022, April 19). How to solve AI's inequality problem. *MIT Technology Review* online. Available here: https://www.technologyreview.com/2022/04/19/1049378/ai-inequality-problem/?gad_source=1&gad_campaignid=16701804390&gbraid=0AAAADgO_mhAnko6NtaB33Igvc67uKGB2&gclid=Cj0KCQjw_b_QBhCSARIsAP6hR4c6ENb83BHL2j9Dh70mmj3011m0zcTvgGTp9GEmliABvnn6gIbsQI0aArsJEALw_wcB

[49] Dubey, A. (2025 June 4). How AI can enhance digital inclusion and fight inequality. World Economic Forum. Available here: https://www.weforum.org/stories/2025/06/digital-inclusion-ai/

[50] United Nations Sustainable Development Goals Report (2025): https://unstats.un.org/sdgs/report/2025/

[51] Linking Artificial Intelligence Principles (LAIP) is a collaborative initiative and platform designed to synthesize, compare, and analyze the various AI ethical guidelines proposed by governments, research institutes, and technology companies worldwide. **However, it is not a secure site**: https://linking-ai-principles.org/.