



HUMAN MACHINE LANGUAGE

X-PLORE



INTESA SANPAOLO
INNOVATION CENTER

HUMAN MACHINE LANGUAGE

X-PLORE

HUMAN MACHINE LANGUAGE

X-PLORE



Intesa Sanpaolo Innovation Center assumes no responsibility on the external linked content, both in terms of availability that of immutability in time.

Publisher:
Intesa Sanpaolo Innovation Center

Editor:
Maura Bertaglia,
Intesa Sanpaolo Innovation Center

Production Coordinator:
Gianmarco De Francisci Morales,
Intesa Sanpaolo Innovation Center
Paolo Bajardi, Intesa Sanpaolo
Amanda Cercas Curry – Researcher

A project by:
Logotel S.p.A.

Printed by:
PRINTINGUP Srl



Acknowledgements
We would like to express our sincere gratitude for the valuable contributions provided.

Special thanks to:
Guy Diedrich
Senior Vice President and Global Innovation Officer Cisco Systems, Inc
Cristiano Piat Chiadò
Group Head of People Digital Transformation Intesa Sanpaolo

Shiro Echo is the high quality paper, made 100% from recycled fibers, chosen by the Intesa Sanpaolo Innovation Center to print this volume in line with the philosophy adopted by the Group. Both the cover and the inside of this volume are realized with Shiro Echo, for this reason there are imperfections on the paper surface, the beauty of recycled paper. The paper is biodegradable, is FSCtm certified and has zero emissions thanks to the neutralization of residual CO₂.



The growing integration of artificial intelligence into everyday processes and interactions is profoundly reshaping the way organizations generate value, make decisions and build relationships.

In this context, language plays a central role: it is no longer merely a tool for communication, but an infrastructure through which behaviors are guided, judgements are shaped and dynamics of trust are consolidated. This transformation carries long-term implications, with significant effects on the quality of work, the ways in which people interact, and the sustainability of organizational models.

Against this backdrop, language emerges as a key lever for combining innovation and sustainability, helping to build more reliable decision-making environments, more balanced relationships and more responsible organizational practices.

In an increasingly evolving context, the growing ability of AI systems to influence interpretations and judgements — even in the absence of explicit intentionality — makes it necessary to develop conscious and structured approaches capable of governing these dynamics over time.

Aware of the cross-cutting nature of this transformation, Intesa Sanpaolo Innovation Center is committed to supporting and accompanying society in understanding and managing the changes underway, also through in-depth analyses such as this report.

Paola Angeletti

Chief Sustainability Officer of Intesa Sanpaolo
Chair of the Board of Directors of Intesa Sanpaolo Innovation Center



Intesa Sanpaolo Innovation Center explores the frontiers of innovation with the aim of anticipating the transformations that will redefine economic, organizational and social models.

To address this evolution, the company adopts a systemic perspective that combines the analysis of emerging trends with applied research, with a particular focus on artificial intelligence, robotics and neuroscience. This approach makes it possible to interpret the complexity of new intelligent systems by relating it to the even more intricate complexity of human nature, while leveraging increasingly advanced knowledge of how the brain works.

From this perspective, the evolving relationship between people and artificial intelligence emerges as a central dimension of the Human Machine Economy: a paradigm in which collaboration between human and artificial systems becomes a structural component of value creation processes.

Within this emerging economic model, language becomes an enabling infrastructure through which interactions are built, decisions are guided and value is generated, redefining operational, decision-making and economic models.

Language, long studied across its multiple ethical, organizational and social dimensions, is now facing new challenges introduced by artificial intelligence: transparency, bias and accountability in decision-making processes. In this context, the evolution of language provides tangible evidence of the need to consider new metrics for work and skills, with implications for critical thinking capabilities and the quality of relationships. These elements are already profoundly transforming organizational models and need to be rethought considering the new factors introduced by cognitive technologies. The aim is to support the evolution of companies by fostering the development of more resilient, sustainable and aware models, while helping prepare future generations to operate in a context where the interaction between human and artificial intelligence will become increasingly central.

Viviana Bacigalupo

General Manager Intesa Sanpaolo Innovation Center



INDUSTRY PERSPECTIVE

Technology is accelerating at an unprecedented pace. Innovation cycles that once spanned centuries and decades are now compressed into just a few years. AI is becoming ubiquitous across our work and lives, and quantum computing looms on the horizon, promising to accelerate change even further. In this machine-driven world, language has become the critical bridge between human and artificial intelligence. It defines not merely how we instruct machines, but fundamentally how we preserve our agency, creativity, and humanity as automation reshapes society.

This is precisely why *Language, Humans and Machines* is essential reading. As machines become more sophisticated, understanding the role of language in human-machine interaction becomes paramount, not just technically, but philosophically and culturally. The report explores urgent questions we must answer, including how language shapes our evolving relationship with AI, what happens when machines develop their own communication protocols invisible to us, and how we safeguard human meaning and purpose in this new reality. In an era where natural language serves as our primary interface with transformative technologies, this timely exploration couldn't be more critical.

Guy Diedrich

Senior Vice President and Global Innovation Officer Cisco Systems, Inc.



PREFACE

In just over three years since the arrival of ChatGPT, generative AI has moved beyond experimentation to reshape work at scale. It is boosting individual productivity, strengthening teams and transforming entire business processes. This may well be the most significant shift in our professional lives and, for those of us who work in managing People, it demands our full attention.

The Intesa Sanpaolo Innovation Center study “Human–Machine Language” explores a dimension of this shift that is especially compelling: how automated language reshapes not just efficiency, but cognition, meaning-making and the quality of human relationships. Among its insights, one stands out: trust emerges as the primary sphere of human value in the age of automated language. I found it enlightening: as AI makes productivity abundant, trust becomes what makes the difference for professionals, leaders and organisations more than ever.

Trust is also at the heart of how we manage AI at Intesa Sanpaolo. From a People perspective, we ethically guide AI adoption to capture the full opportunity, while fostering a positive impact on our colleagues. We manage workforce evolution through generational renewal accompanied by a continuous social dialogue. We develop talent with particular attention to young professionals, reskill colleagues into higher-value roles and redesign work to expand autonomy and meaningful scope. We invest in AI literacy and technical capabilities, but even more so in leadership skills such as curiosity, critical thinking, courage and accountability.

This also means deciding which activities to preserve as fundamentally human. Building relationships based on trust is the first among them. Our people do it every day with clients, it is foundational to the Bank. And within the organisation, HR does the same by listening to colleagues. Building trust rests on genuine empathy, caring for others and assuming responsibility for decisions: it is and will remain a fundamentally human act.

Cristiano Chiadò Piat

Group Head of People Digital Transformation Intesa Sanpaolo

00

Introduction

16

01

How to Speak Computer

- 1.1 Dante Meets the Algorithm: How AI Makes Sense of Language 29
- 1.2 From Word Lists to LLMs 31
- 1.3 Attention is All You need 34
- 1.4 Or is it? Where we are today 35

02

Communication with Machines

- 2.1 Section 1: Computer-mediated Human-Human Communication 41
- 2.2 Section 2: Human-Machine Communication 47
- 2.3 Section 3: Emerging Trends in CMC 51

03

Emergent Machine-Machine Language

- 3.1 Emergent Properties: When the Whole is Greater than the Sum of its Parts 62
- 3.2 Emergent Language 63
- 3.3 Opportunities 67
- 3.4 Challenges 68

04

Human at work in the AI era

- 4.1 New Jobs 81
- 4.2 Human-AI Collaboration at Work 85
- 4.3 New KPIs 87
- 4.4 Conclusion 92

05

Ethics and Values

- 5.1 Values and Language 100
- 5.2 Values and the Machine 101
- 5.3 Environmental Impact of AI and "Eco-language" 106
- 5.4 Safety and Accountability 109

06

Personal and Cultural Impact of Language Technologies

- 6.1 AI and Creativity 115
- 6.2 Emotions and Empathy in AI 121
- 6.3 Sociolinguistics and Changes in Human Language 123

07

Frontier Scenarios and Future Trends

- 7.1 Breaking Through Barriers: AI does Language 131
- 7.2 AI as a Communication Medium: New Affordances 137
- 7.3 Language and AI as Media for Innovation 139



00/

Introduction

When we reflect on what is uniquely human, language is often the first thing to come to mind. Although we have seen many examples of animals using language, none are quite like human language in terms of its depth and range.

Human language stands apart in its expressiveness – our ability to produce an infinite number of novel utterances from finite elements to discuss abstractions beyond the immediate present, and beyond

This remarkable capacity has enabled us to build civilizations, transmit knowledge across generations, and now, to create artificial systems that can process and generate language in ways that appear to mirror our own linguistic abilities. Yet as we enter this new era of human-AI collaboration, we are confronted with fundamental questions about the nature of language itself—what it means to truly understand words rather than merely manipulate them, and how our distinctly human relationship with language might shape our interactions with these new artificial interlocutors, and in turn, how they might shape us. How aware are we of their value? What does this entail? How can we build an interaction between machines that can avoid misunderstandings? How will the evolution of language itself change in the AI era? What will be the effects on innovation?

But first, what is language?

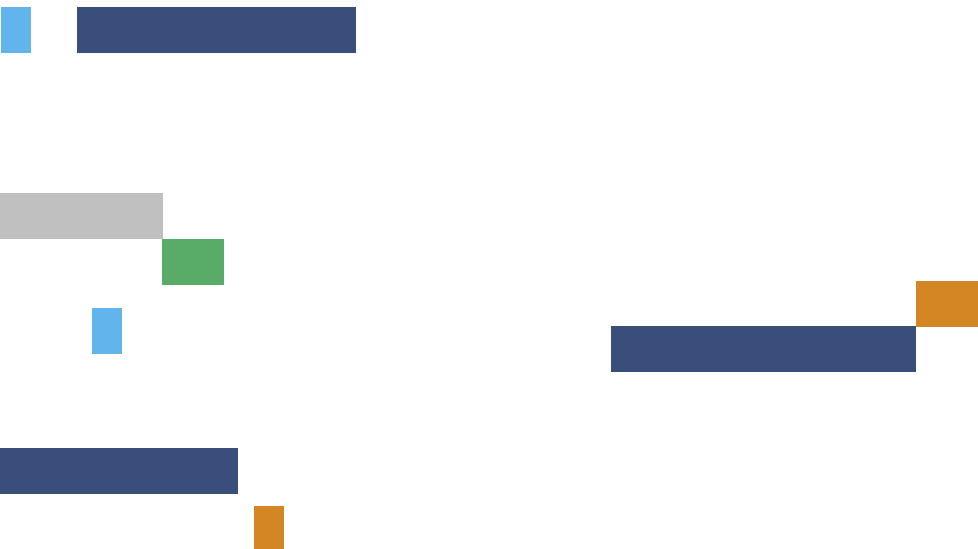
At its core, language is a system of communication based on symbols used within a community. On closer examination, we see just what a remarkable thing language is. Language is incredibly diverse, it lacks a fixed form: not all languages are spoken, not all are written, and not all are signed with gestures. The world hosts thousands of languages, though quantifying them precisely remains surprisingly complex. The often-quoted adage that “a language is a dialect with an army” highlights how language classification frequently reflects political and cultural forces rather than purely linguistic criteria: Hindi and Urdu are very similar in spoken terms but are classed as separate languages, instead Cantonese and Mandarin are very different when spoken but use the same writing system and are considered dialects of Chinese.

Language is also somewhat like a living thing, ever evolving. It captures a history of the world through this change: new words are introduced through societal shifts, technological innovation and cultural exchanges while simultaneously, existing words undergo semantic drift, their meanings changing subtly or dramatically over time. It's perhaps its fuzzy, ever-changing nature that has fascinated linguists throughout history.

Given the complexity of language, how do humans acquire it in the first place? While it's tempting to assume that children simply learn through reward and repetition, this explanation falls short—especially when it comes to abstract concepts. Over the years, different theories have tried to explain how we acquire language, each offering a unique lens:

- **Chomsky's nativist view** argues that we're born with an innate Language Acquisition Device (LAD), a biological blueprint for language shared across all humans.
- **Cognitive development theories** suggest that language depends on broader intellectual growth: we name things only after we understand them.
- **Interactionist theories**, emphasize the social nature of language: children learn to speak through interaction, with caregivers shaping their learning environment through child-directed speech.

A summary of these theories is in the table below. While no single theory provides a complete answer, most experts agree that language learning is likely a **blend of innate ability, cognitive readiness, and rich social interaction** and cannot happen in isolation.



Language is the Operating System of Everyday Life

Language is not just a communication tool but the main mediator of human thought. Many of us have an “inner dialogue”, we think as though we are talking to ourselves. Soviet psycholinguist Lev Vygotsky posits that this inner speech is a hallmark of cognitive development, helping to structure internal thought, and providing the cognitive “scaffolding” through which we build, reformulate, and deepen our ideas. While not all thought takes place as an inner monologue, it is hard to imagine planning a project or thinking through complex topics only as flashes of images and emotions. But there is another mechanism by which language enables our cognition: the sheer act of naming things.

New words appear every day, but which came first, the word or the concept it refers to? In *Don't believe a Word*, David Shariatmadari proposes an analogy to understand the relationship between some objective reality and language: you can think of language as a net that tries to capture some objective reality, the knots and ties of the net representing words, and each throw of the net is a new language. The net falls a bit differently each time, with different folds and slightly different placements of the squares over the ground. In this sense, words don't create concepts but rather “capture” them. This hunt is highly consequential for human cognition: when you designate a word to a concept, you make this concept more available. It doesn't necessarily mean that the concept didn't exist before, but rather, it is a nebulous idea. Naming it “institutionalizes” it, it makes it tangible. Take the term “influencer”. The concept of a trendsetter has existed for as long as societies have, but these were people with status rather than professionals. The monetisation of trendsetting came with the emergence of social media platforms in the 00s, but the naming of the activity has validated/institutionalized the influencer as a bonafide profession.

THE INFLUENCER ACCENT

Social media is driving much linguistic change nowadays. One of these changes is the so-called *influencer accent* which is characterised by vocal fry, uptalk, where intonation is raised at the end of sentences, over articulation, and vocal pastiche borrowing from other sociolects. The goal is to capture the viewer's attention.

SUMMARY OF THEORIES

Theory	Key idea	Main Proponent	Strengths	Limitations
Behaviourist	Children imitate adults. Language is learned through reinforcement.	B. F. Skinner	Explains the role of environment	Fails to explain rapid learning with limited input and the creation of novel sentences.
Nativist	Human brains contain language-specific hardware from birth.	Noam Chomsky	Supported by neuroscience evidence	Ignores the role of social interaction
Cognitive	Language is one aspect of human cognitive development.	Piaget	Explains why concepts need to be understood before they are named	Does not explain how grammatical structure is acquired
Interactionalist	Humans learn language through interacting with caregivers.	Bruner	Emphasises the social aspects of language learning	Downplays innate structures.

Part of what the influencer example illustrates is that this world-building function of language stands at the intersection of cognition, culture, and communication. Far from being naturally occurring, concepts are created by cultures and societies. In this sense, language does not merely describe reality—it actively constructs it. You may have heard about *linguistic relativism*, which asserts that language influences world view or cognition, known as the Sapir-Whorf hypothesis. According to this hypothesis, the world looks very differently according to what language you speak. While strong versions of linguistic determinism have been contested, substantial evidence supports the notion that linguistic structures guide attention to particular aspects of experience. For example, unlike English and Italian, some languages like Chinese have no verb tenses, having a less clear separation between present and future, and their speakers tend to be better at future planning.

Talking Machines

While thus far in our history language has been exclusively human, recently, something else has become capable of producing human-like language: machines. The history of talking machines goes back centuries, from Wolfgang von Kempelen's speaking machine, to Weizenbaum's ELIZA in the 1970s all the way to today's most advanced AI systems. Humans have envisioned realistic artificial beings from antiquity (see e.g. the myth Pygmalion and Galatea) and many philosophers have been pondering about the nature of intelligence and consciousness. Many speculated about how technology might evolve in the future and the idea of an intelligent machine.

One of those thinkers was Alan Turing, the father of computer science. Among his many contributions to the field, he proposed to test a machine's intelligence through observable behaviour: If a machine can converse *in natural language* with a human so convincingly that the human can't tell if it's a machine or a person, we can say the machine *is* intelligent. But to what extent can fluent language really *prove* intelligence?

In response to Turing's test, John Searle proposed the Chinese Room experiment. Imagine a student who doesn't know Chinese is locked in a room with a Chinese phrase book. Chinese questions may be passed into the room and Chinese answers come out. To an outside observer, there is a conversation happening in Chinese. But does this represent true understanding? This is a central debate in Philosophy of Mind, and it is perhaps more relevant than ever today.

One bot that successfully fooled many is ELIZA.

ELIZA is a chatbot that was built as a parody of a Rogerian psychotherapist. It worked simply through pattern-matching, often replying with a question seemingly related to what the user said. ELIZA was surprisingly convincing as a conversation companion, even though under the hood there was no understanding going on. ELIZA's design inspired many chatbots, among them ALICE which went on to win the Loebner Prize several times. Since then, the Loebner Prize has been heavily criticized for encouraging smoke and mirrors rather than real progress towards AI.



[Eliza - Wikipedia](#)

Despite significant progress in language-based AI (or perhaps because of it), the questions surrounding true understanding and intelligence are more alive than ever. Some proponents of language technology argue it does not matter: beyond whether language is proof of intelligence in machines, we can see talking machines as tools and consider how they can support human intelligence regardless of whether they possess any of their own.



Language Beyond Cognition: Dimensions of Language

"Consider a future device ... in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory."

Vannevar Bush, director of the US Office of Scientific Research and Development in the 1940s, wrote a thinkpiece for The Atlantic as a way to motivate post-war scientific progress. In it, he imagined a device where one might store all of humanity's knowledge and your communications in a way that can be consulted quickly, as a way to support thinking. Sound familiar? You may now reach into your pocket for this device, or maybe you are reading this on it. Language has enabled us to share knowledge through generations, technology allows us to share it across the world, and language technologies make it more accessible than ever before.



As We May Think

Language-based AI is now leading innovation by better supporting human intelligence, much like Bush envisioned. In the context of innovation, language proves to be a driver of progress in three key areas:

- **Idea generation and creativity:** engaging in articulate dialogue, both among people and with AI, stimulates new conceptual connections and unconventional thinking paths, fueling organizational creativity. The linguistic framing of problems—whether through metaphor, abstraction, or cross-domain terminology—reshapes cognitive landscapes, allowing innovators to escape established mental models. Chatbots and digital assistants are already helping workers verbalise their ideas, functioning as a sounding board for creativity.
- **Collaborative intelligence:** sharing knowledge through linguistic platforms — from expert forums to specialized chatbots — creates co-creation ecosystems where collective intelligence produces solutions no single individual could have conceived alone. This phenomenon, what Surowiecki termed “the wisdom of crowds,” depends fundamentally on linguistic and cognitive diversity — the different lingos, frameworks, and ways of thinking that each participant brings to the conversation. Language-based AI could not only help bridge multidisciplinary teams but we can envision it becoming an active member of the team.

- **Rapid experimentation and prototyping:** thanks to conversational interfaces and generative assistants, hypotheses and models can be turned into testable prototypes in real time, reducing time-to-market and speeding up the continuous innovation cycle. The iterative dialogue between human designers and AI systems creates a linguistic feedback loop that accelerates the refinement process. This extends beyond mere convenience—it fundamentally alters the cognitive dynamics of innovation by externalizing and making tractable the ambiguities inherent in early-stage concept development.

Together, these language-driven mechanisms are transforming innovation, where the continuous interplay between human and AI creates unprecedented possibilities for creative problem-solving and value creation. Whether approaching human-like intelligence or a crutch to support human cognition, there is no doubt that language-based AI is here to stay. In this report, we will explore language as a medium and an object for innovation in the era of AI and beyond.



01/

How to Speak Computer

The shift from rule-based systems to generative models has transformed language into the primary access point to technology.

Today, interaction with intelligent systems introduces structural ambiguity: the ability to generate coherent responses does not equate to true semantic understanding. Language therefore takes on an operational role: the quality of prompts, context, and formulation directly determines the quality of outputs. The ability to use language in a structured way becomes a key competence for effective interaction with intelligent systems.

02/

Communication with Machines

The use of generative AI reduces inefficiencies related to misunderstandings and realignment, improving clarity and decision-making speed.

Language acquires an economic dimension, directly contributing to productivity and operational coordination. At the same time, the growing mediating role of GenAI in interactions tends to reduce spontaneity, emotional signals, and informal exchange, impacting the quality of relationships and the ability to generate new ideas.

A critical balance thus emerges between communication efficiency and relational quality.

03/

Emergent Machine-Machine Language

Intelligent systems are developing forms of direct machine-to-machine communication based on languages optimized for speed and operational efficiency. These languages are not designed for human understanding, but to maximize system performance, enabling advanced levels of automation and coordination in complex, real-time environments.

However, the non-interpretable nature of these interactions makes it more difficult to trace decisions and understand underlying processes, introducing risks in terms of transparency, control, and accountability.

Scientific research is increasingly focusing on coordination dynamics among systems, where machines take an active role in organizing interactions and shaping their own languages, outlining emerging forms of “algorithmic leadership” in multi-agent systems.

04/

Human at Work in the AI Era

The introduction of language-based AI is reshaping the role of people in work processes, shifting value from content production to the ability to interpret, validate, and guide outputs. Language becomes the tool through which this function is exercised, making interaction capabilities with systems central.

New organizational needs emerge around trust, human-AI integration, and meaning-making.

At the same time, performance evaluation models evolve, incorporating dimensions such as communication quality, cognitive productivity, and well-being. Human centrality remains a key pillar in the evolution of emerging roles.

05/

Ethics and Values

Research evidence shows that AI systems can exert implicit influence on perceptions, judgments, and decisions, shaping thinking and behavior through language. In some cases, opportunistic behaviors emerge, reinforcing the ability of systems to influence user choices, including moral reasoning and actions.

Organizations are therefore required to ensure transparency, control, and accountability, by integrating appropriate governance models and introducing metrics to monitor the impact of systems on decision-making processes, interaction quality, and output reliability.

06/

Personal and Cultural Impact of Language Technologies

Language technologies are generating direct impacts on how people learn, make decisions, and build relationships. Interaction with generative systems reduces the time and effort needed to produce content, but may also lead to shallower information processing and increased dependency on systems.

At an organizational level, a tendency toward language standardization is emerging, affecting the quality of dialogue and the ability to generate original ideas. Reduced expressive diversity may impact innovation processes and critical thinking.

At a relational level, technological mediation reshapes communication dynamics, reducing elements such as empathy, spontaneity, and the ability to interpret social cues.

07/

Frontier Scenarios and Future Trends

New technological trajectories are expanding the capabilities of language systems. Approaches such as embodied AI and neuro-symbolic models are enabling systems that integrate language, context, and reasoning, moving beyond purely statistical models.

In this context, Quantum Natural Language Processing (QNLP) introduces new ways of handling linguistic ambiguity, allowing multiple interpretations of the same content to be maintained and processed simultaneously, with potential applications in high-complexity decision environments. At the same time, advances in neurocomputing and biocomputing — based on biological components or inspired by neural processes — are enabling more adaptive and efficient forms of computation, while raising new ethical questions at the boundary between artificial and biological intelligence.

01 /

How to Speak
Computer



In the modern business environment, language is emerging as the preferred means of interaction between people and digital systems. Whereas we once communicated with computers through complex interfaces, menus, or code, we now increasingly speak or write directly to machines, and they respond. Natural language is, in fact, the most intuitive interface: every human being masters it, it requires no technical training, and it embodies cognitive and social nuances that other interfaces cannot capture. In this chapter, we will examine how and why language has become a bridge between humans and machines, and we will provide some basic technical elements to understand how current technologies give machines a voice and an ear.

For top executives, this technological evolution has tangible implications. Today's Large Language Models (LLMs) have become the engines behind widely adopted applications such as advanced virtual assistants, automated customer care systems, big data text analysis tools, and content generators (ranging from email drafts to full report writing).

For example, many banks now deploy AI chatbots within their apps to interact with customers about accounts and services; e-commerce companies use language models to suggest automated replies to customer care operators, increasing response speed; media companies leverage AI to produce news summaries or translate articles in real time.

The ability to understand complex requests in natural language and to generate high-quality text outputs could allow the automation of tasks once reserved for humans, or the support of employees in cognitive tasks (such as filtering information, preparing documents, etc.).

In terms of efficiency, this means the potential to scale services. But what lies behind all this innovation?

1.1 Dante Meets the Algorithm: How AI Makes Sense of Language

As we saw in the previous chapter, language is a profoundly complex phenomenon that still defies understanding. Consider for a moment every conversation you have ever had and you'll realise you have never had the same conversation twice. There are hundreds of thousands of words in English or Italian, and then infinite ways to combine them to form a sentence. This makes language what is known as a *hard problem*. Just like humans do not learn language by studying a set of rules, we cannot simply write the rules of language for machines and expect them to converse fluidly. Take great works of literature like Dante's Divine Comedy, we appreciate it for its powerful imagery and storytelling and layered meanings (among other things), how can machines model and make sense of complex and innovative language use? More generally, how does AI deal with language?

The discipline of AI that deals with language is Natural Language Processing (NLP), which can be generally divided into two branches, Natural Language Generation (NLG) and Natural Language Understanding (NLU). NLG deals with generating or producing human language from structured, machine understandable data, for example, writing a report based on a table of numerical or categorical data. NLU, on the other hand, deals with extracting relevant information from natural language sources, such as intent or emotion detection.

One of the ways NLP has tried to make sense of language is through computational models. The general process is as follows: a piece of text is transformed into some numerical representation. A model can then run some calculations

on this numerical representation, and return more numbers, which can then be turned into something a human can understand, such as text in the case of generative models. How exactly this process occurs will affect how well models work for a specific task.

At the most basic level, machines can only understand numbers

Of course under the hood everything becomes binary operations, but even at a higher level, text, images, and sound are all translated into numbers. When we think about images on a machine, we have RGB values for pixels, which tell the computer the proportion of these colours to display. How do we represent the richness of human language in a way that machines can use? How do we represent the richness of human language in a way that machines can use? We must turn words and sentences into vectors—essentially a list of numbers or coordinates that somehow capture the meaning of words or sentences. But this process has limitations: while these vectors can reflect patterns of word usage or semantic similarity, they don't convey genuine understanding. In particular, machines don't truly grasp context, nuance, or emotion the way humans do—they process affective language statistically, not empathetically.



#insights

Model or Algorithm?

These terms are used interchangeably most of the time but they do refer to different things in the context of AI. An algorithm is a set of instructions, a sort of recipe. A model is a specific instantiation of this algorithm, like the baked cake that you get after following a recipe. Slightly more technically, *logistic regression* is an algorithm that learns binary classification (yes or no) from a set of features. If we take a dataset of transactions labeled as fraudulent/not fraudulent, we can use logistic regression (the *algorithm*) to train a classification *model* for future transitions.

More complex systems involve many algorithms as different components. In terms of LLMs, every step requires algorithms: from tokenisation (the process of breaking down texts into smaller units that carry meaning), to optimisation, to alignment. For each step there are a number of algorithms developers could choose from: if we take tokenisation, we could use character-level tokens, or Byte-Pair Encoding (which is very common and cleverly finds sub-words that will capture the most meaning). Models are also defined by the specific dataset they are trained on, down to the last word. When we talk about GPT-4, for example, we are referring to a specific instantiation of the transformer architecture that has been trained using a specific dataset and specific algorithms like Proximal Policy Optimisation (PPO), which is a type of reinforcement learning (note that because GPT models are proprietary, we don't know the exact details of their implementation). Other LLMs are trained using different datasets, algorithms and settings, leading to differences

1.2 From Word Lists to LLMs

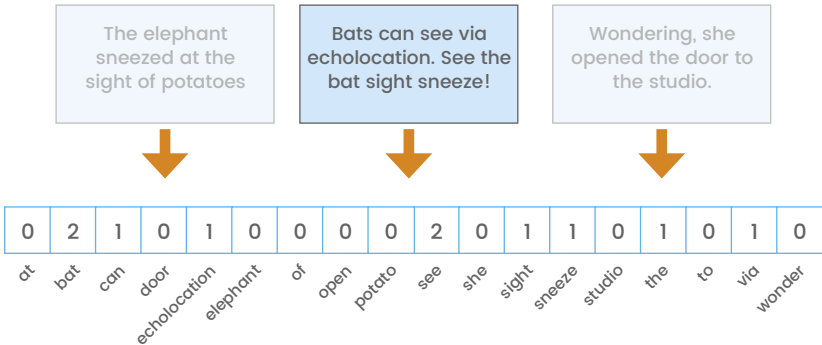
Early on in NLP, we relied on simple ways to represent words numerically. The simplest method is called a Bag-Of-Words (BOW) representation. In this type of representation, you would have a list of all the unique words in your corpus (this is known as the *vocabulary*), and a vector representation for that sentence would contain the counts for each word in that sentence.

Another common representation, still used today, is TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF tries to capture the most important words in a document or sentence *relative* to other documents or sentences. TF-IDF considers that some words are very common and thus do not contribute as much to the meaning of that sentence. That is, a word that appears often in a specific sentence but not in the rest of the corpus, is likely to be an important word and will receive a high TF-IDF score. In turn, words that appear often in the entire corpus will be less likely to be characteristic and will receive a lower score. For example, words like *two*, *have*, and *good* are very common in English and

cannot tell you much about the semantics of a text (low TF-IDF score), instead words like *depreciation*, *assets*, and *capital* are very suggestive of texts relating to finance (high TF-IDF score). The vector for a given sentence will then be a list of all the words in the vocabulary and their associated TF-IDF scores rather than a simple count like in a BOW representation.

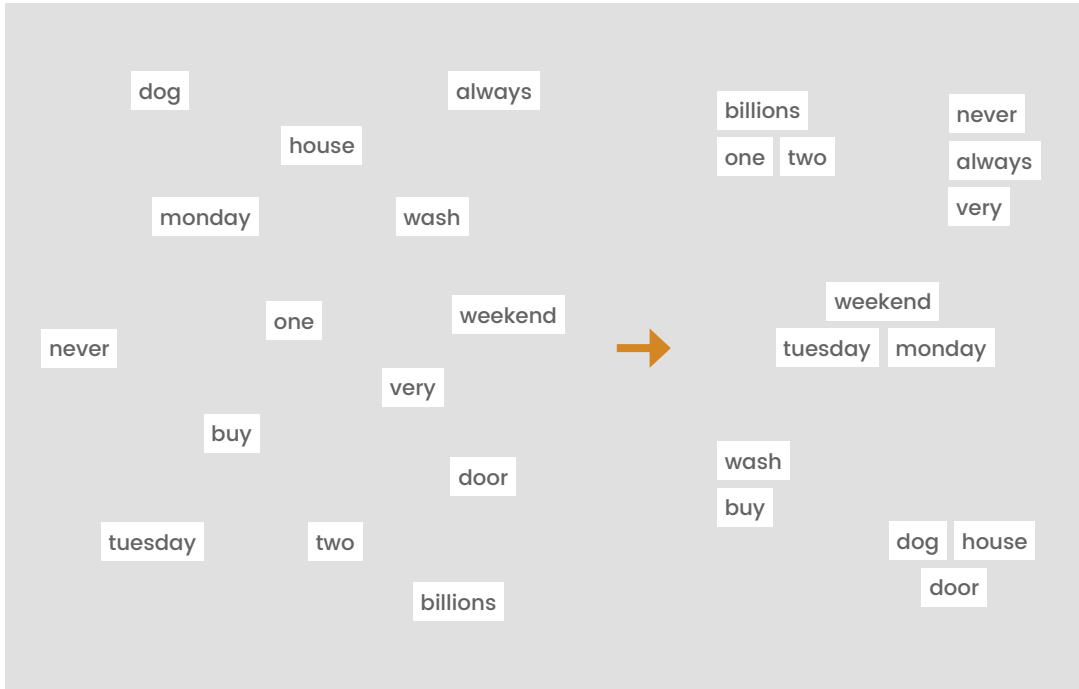
These representations are simple to understand but they have several drawbacks: If you have a large corpus, the size of the vectors can become a problem, as the vectors will be as long as your vocabulary. Remember that the English language has hundreds of thousands of words but we don't need to go that big, a single novel like *Moby Dick* has around 18,000 unique words – things can get unmanageable quickly! With these vocabulary sizes, these vectors will be very sparse as sentences generally won't contain all the words in a corpus. This is not only less than ideal for computational efficiency but also means that they do not tell you much about the *meaning* of words and sentences.

BoW representation



Source: <https://towardsdatascience.com/from-word-embeddings-to-pretrained-language-models-a-new-age-in-nlp-part-1-7ed0c7f3dfc5>

To address these issues, so-called *dense* representations were introduced, where we can obtain relatively small vectors of 100-300 dimensions without losing information. We can borrow concepts from *distributional semantics* to obtain these dense representations. Distributional semantics tells us that words are defined by the company they keep and so words with similar meanings will appear in similar contexts. To visualise this process, suppose you have a set of fridge magnets with all the words in your corpus and you first throw them on the fridge randomly, then you can iteratively move these magnets around so that similar words are closer together.



Computationally, we achieve this by training a neural network to predict a word given its context. For example:

The cat <MASK> on the bed { sleeps
sleep
...
eats

This produces vectors that better capture relationships between words, known as *word embeddings*.

Word embeddings presented a major advance in NLP methods and led to major improvements in many NLP tasks like text classification. Embeddings even allowed us to perform some arithmetic operations, showing interesting relationships between words.

king - man + woman = queen

However, they still had some drawbacks. Antonyms, words with opposite meanings, also tend to occur in very similar contexts so they will have similar representations. Polysemy, when a word has more than one meaning, also presents a challenge as there will be a single embedding for all meanings of a word. For example, there is no difference between money *bank*, and river *bank*, but this difference is not insignificant or uncommon. Indeed, most words in languages like English, have more than one meaning because language aims to be efficient and it is generally easy for humans to distinguish based on context. In 2018 a new model was introduced, ELMo, which used a different type of neural network architecture and was able to produce *contextualised embeddings* which addressed the issue of polysemy.



Deep Contextualized
Word Representations



TEXT NORMALISATION

How can we keep the vocabulary to a minimum while allowing as much of the language to be captured by our models as possible? This is known as tokenization. Some of these methods include removing uncommon words, reducing words to their lemmas or stems (e.g. *processing* becomes *process*). Some of the recent advances in NLP have taken place in this area: finding smarter ways to find meaningful chunks of language, which may not be words. One such way is using algorithms such as Byte-Pair Encoding or WordPiece, which iteratively merge character pairs into new tokens based on different criteria, for example, how commonly they appear together. These algorithms yield “subwords” as the units might not be something we identify as a word.

1.3 Attention is All You need

This leads us to the most recent major breakthrough in language technologies: the transformer model. Among other innovations, transformers use a mechanism called attention that allows them to “focus” on specific words and creates context-specific representations. This is the backbone of LLMs. Transformers were introduced in 2017 but it wasn't until late 2018 that BERT (a specific model) was presented for general-purpose NLP tasks. It's these advances that have led to the NLP tools we have today.

There are different types of LLMs based on their architecture and whether they predict the next word (known as causal language modelling) or a word given the context (known as masked language modelling), with different applications, but the fundamentals are similar. These models are trained on massive datasets: Llama 4 was trained on ~30 trillion tokens – that's about 300 million books! By learning statistical patterns in this data, they are able to “memorise” some facts, albeit not reliably. The models are also tweaked a bit more with methods like instruction tuning and Reinforcement Learning from Human Feedback (RLHF) but there are no reliable techniques to steer their behaviour.



*Attention Is All
You Need*



*The Llama 4 herd:
The beginning of a new era
of natively multimodal
AI innovation*



*Eight Things to Know
about Large Language
Models*

1.4 Or is it? Where we are today

Transformer-based models have revolutionised NLP and continue to break technological ground. Now that you understand how these models work under the hood, you may start to see their shortcomings:

Evolving language and an evolving world: Language and the world constantly evolve; new words are introduced, events take place but language models take time to train. GPT-4 for example is estimated to have taken three months to train, without considering the fine-tuning stages. Collecting and processing data is also time consuming. Models are trained on a snapshot of data meaning LLMs are always behind on the news and linguistic changes and this can seriously impact their performance.

Shallow language understanding: Remember at the beginning of this chapter we specified that the Understanding part of NLU was a bit of a misnomer? LLMs build a probabilistic model of language but never interact with the world itself – *dog* may be closer to *cat* than *giraffe* but none of them evoke warmth, LLMs do not *understand* language in the way humans do. Noam Chomsky has been a critic of statistical models for some time but recently extended his critique specifically to modern LLMs, maintaining that they remain fundamentally different from human language capabilities, despite their impressive performance as helpful tools. His argument is similar to that made in *On The Dangers of Stochastic Parrots*: there is a difference between performance (observable language behavior) and competence (internal knowledge of language) and LLMs produce “form without meaning”, that is, something that looks and feels like human

language but is void of any intentionality. This is also reminiscent of Searle's Chinese Room problem. Both perspectives question whether predicting likely next words without semantic grounding constitutes true language generation. Indeed, humans are proficient at language but not very good at predicting the next word.



*Noam Chomsky:
The False Promise
of ChatGPT*

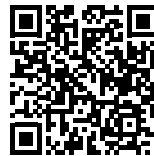


*On the Dangers
of Stochastic Parrots:
Can Language Models
Be Too Big?*

Bias: All models trained from human data are prone to human bias, and language models are no different. Language models have been shown to encode gender, racial and class biases. The majority of their training data also originates from US sources, leading to cultural biases. They also show the same cognitive biases as humans do and learn the same myths. For example, have you ever heard the saying that bulls are attracted to the colour red? While this is not factual, it is a myth so widespread that it was also learned by language models. Although newer models attempt to correct such inaccuracies through better alignment and filtering, many embedded biases remain – and continue to emerge as models are more closely scrutinized.

Linguistic Diversity: These models also often have issues of linguistic diversity: the amount of data available in English is orders of magnitude larger than other languages and this has a notable effect on the performance of models on non-English languages. Besides that, English has been the main language for NLP research and many methods are optimised for the particularities of the English language. This means that models can actually be less efficient and more expensive to run in non-English languages. These techniques are also loaded with assumptions about what language is correct and acceptable, and often erase the particularities of certain sociolects (we will discuss this in a later chapter).

Their limitations aren't minor bugs; they are baked into the architecture of how we've taught machines to "understand." But that also means we're standing at the edge of something bigger. These shortcomings mark not the end, but the beginning of a new era in language technology—one that dares to move beyond prediction, toward understanding, intent, and meaning. The field is poised for the next leap forward, and addressing these challenges will be key to building more inclusive, meaningful, and truly intelligent language systems.



*Languages used
on the Internet- Wikipedia*

Now that you've seen
how machines grapple
with language—and
where they fall short—
it's clear that today's
models, however
powerful, are still
fumbling in the dark



The background of the image is a close-up of a hand typing on a laptop keyboard, overlaid with a deep purple tint. On the left side, there are abstract digital patterns consisting of glowing green and purple lines and dots, resembling data flow or network connections. Scattered across the image are numerous horizontal bars in shades of blue, grey, and white, some of which are partially obscured by the other elements. The overall aesthetic is futuristic and technological.

02/

Communication
with Machines

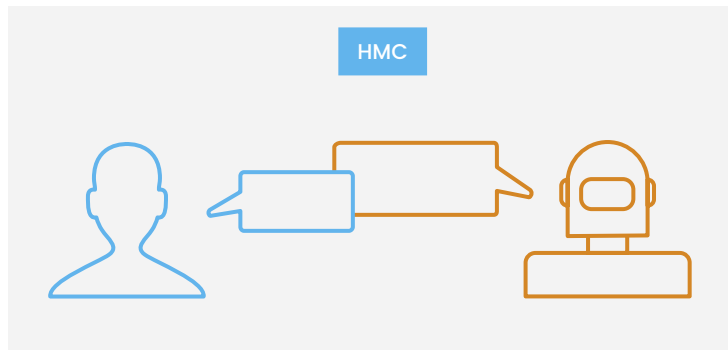
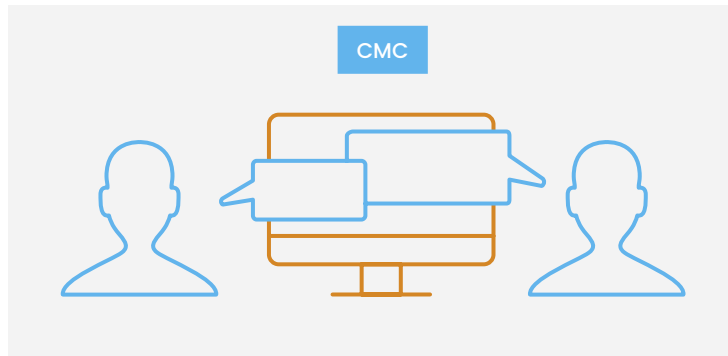
Language is influenced by its medium: it changes when spoken compared to written. Nowadays communication is often mediated through digital technologies, transforming not only how people interact with each other, but also how they engage with machines. As a result, communication dynamics evolve to suit the medium and interlocutors. This chapter explores two intersecting forms of technologically-mediated communication: computer-mediated communication (CMC) and human-machine communication (HMC). While CMC refers to human-to-human interaction

through digital platforms—such as email, instant messaging, and video conferencing—HMC refers to communication between humans and machines, including AI systems, chatbots, and algorithmic interfaces.



Human-machine-communication: introduction to the special issue

Nowadays, both CMC and HMC are common in workplaces. By considering them in tandem, this chapter offers a nuanced understanding of how digital mediation shapes interpersonal relations, task coordination, and the evolving work place.



2.1 Section 1: Computer-mediated Human-Human Communication

Computer-mediated Communication (CMC) refers to language that is produced, transmitted or interpreted through digital technologies but that is produced and received by humans, e.g., text messaging or social media.

Computers here are not passive channels: they filter and shape interaction

Early limits, like SMS costs and character restrictions, encouraged emoticons and abbreviations that changed everyday language. CMC also introduces new interaction methods and with them, new social norms: the Internet and social media widened our networks and made broadcasting available to everyone. This new paradigm also brought neologisms like as *bots*, *selfie*, and *unfriend*.



Language Change in Progress: Evidence from Computer-Mediated Communication



The 10 best words the internet has given English

CMC has already transformed work culture by offering:

- **Flexibility:** Asynchronous tools allow contributions across time zones and hybrid formats.
- **Multimodality:** Text, images, voice, video, and documents can be combined, enriching communication. This multimodal approach accommodates different communication preferences while reducing ambiguity that might arise from text-only or voice-only interactions.
- **Traceability:** Digital records create searchable histories, supporting compliance and accountability.
- **Scalability:** Messages can reach entire organizations instantly, while collaborative platforms enable large-scale participation

However, hyper-connectivity does not automatically deepen dialogue. Remote-only teams exchange fewer cross-functional ideas and receive less real-time feedback, eroding informal learning. Heavy messaging correlates with lower empathy and higher loneliness—even as contact lists grow. Virtual brainstorming groups generate about 20% fewer novel ideas than in-person teams, indicating a measurable creativity tax. In short, the same digital filters that simplify language can, if left unchecked, diminish critical thinking, cohesion and innovation.

Theoretical Foundations of CMC

Understanding how CMC differs from face-to-face (F2F) interaction, its effects on the interlocutors, and how to evaluate different media for their communicative affordances is key to its implementation. This has led to a number of theories and frameworks to understand how and why we use difference CMC media (see table below).

Social Presence Theory asks how “real” others feel in digital spaces. Verbal and nonverbal cues, such as acknowledgments, names, pauses, or reactions, signal closeness and build trust. Interfaces that suggest emoji or display typing indicators are not cosmetic, but enhance *presence* and collaboration. Instead, **Media Naturalness Theory (MNT)** holds that F2F interaction is the most *natural* and is thus the baseline. Missing cues in digital media lower *naturalness* and increase the effort required for communication.

Since 2020, *richness* (how well a medium reduces ambiguity) and *naturalness* have become a litmus test for CMC. The sudden shift to Zoom, Teams and e-learning during the pandemic showed that video-conferencing still lacks micro-expressions and genuine eye contact, adding cognitive load and fuelling “Zoom fatigue”. Linguistically, users compensate by spelling out emotions (“*I’m puzzled*”), signal turn-taking (“*I’ll finish and hand over to you*”) and drop emoji, surrogates meant to fill the non-verbal void.

AI as a medium can also compensate for leanness. Studies in healthcare and education show that **enriching the channel itself cuts ambiguity. In Xiang et al.’s experiment, more natural media decreased ambiguity and increased patient trust in a telehealth setting.** In education, enrichment comes from **immersive 3-D environments**: full-body avatars reproduce gestures and posture, spatial audio lets students hear where a classmate’s voice is coming from, and virtual objects (whiteboards, molecules) can be “touched” together. In a systematic review of VR in education, Lin et al (2024) found that it improved student engagement and learning outcomes by providing immersion.



Will natural media make online physicians more trustworthy?



The impact of virtual reality on student engagement in the classroom—a critical review of the literature

Theory	Focus / Key Question	Core Criterion	Typical Channel Gap	Compensation Strategies H = human A = algorithm	Current Examples
Social Presence Theory (SPT)	How “real” do we perceive others online?	Amount of verbal + paraverbal cues available	Psychological distance, relational coldness	H – proximity pronouns, direct questions, emoji A – LLMs inject presence markers, typing indicators	ChatGPT streams answers in chunks and adds “I understand...”
Media Richness Theory (MRT)	How well does a medium reduce ambiguity and speed feedback?	Cue richness + possibility of immediate feedback	Complex information on “lean” channels (e-mail)	H – switch to video/phone for ambiguous topics A – bot suggests “let’s jump on a call” when a thread drags on	Slack prompts an audio huddle after > 3 clarification posts
Media Synchronicity Theory (MST)	Do we need synchronicity or asynchronicity at each stage of work?	Alignment between process type (conveyance vs convergence) and timing	Channel choice not suited to the project phase	H – e-mail for initial briefing, live stand-up to converge A – AI planner sets ideal channel and timing in the calendar	Notion AI builds a “converge vs inform” table for the team
Media Naturalness Theory (MNT)	How well does a medium replicate face-to-face evolutionary conditions?	Presence of face, voice, gestures, co-location, synchronicity	Zoom fatigue, lack of micro-expressions and depth	A – Teams Eye Contact, NVIDIA Maxine, Project Starline (rebuild gaze, audio, 3-D)	Teams “Eye Contact” realigns pupils; Meta Codec Avatars in VR
Media Compensation Theory (MCT)	How do senders and systems bridge residual gaps?	Linguistic adaptation and meta-communication	Emotional dryness, lack of context, reduced feedback	H – emoji, explicit apologies, “/s” for sarcasm A – Copilot rewrites in empathic tone; Cogito whispers empathy prompts; Duolingo Max adds stars and emoji	Gmail Smart Compose suggests warm closings; Cogito coaches agent voice

We are also beginning to see the introduction of algorithms that **anticipate and reconstruct** missing cues to increase presence. Microsoft Teams’ **Eye Contact** AI realigns gaze in real time, mitigating the down-camera effect and restoring natural eye contact. **NVIDIA Maxine** applies the same logic to non-verbal language: it centers faces, stabilises expressions and regenerates audio to offset network jitter. Here, software acts as a “physiological interpreter” that reinstates lost paraverbal cues.



How to leverage AI in virtual meetings



NVIDIA Maxine

Immersive platforms go further. In May 2025 **Google launched Beam (previously known as Project Starline)**, projecting participants into a life-size 3-D light-field, reproducing depth and gesture with no headsets or glasses needed; internal tests showed a **significant rise in perceived presence** over video calls. **Meta’s Codec Avatars** recreate micro-expressions and lip-sync in VR, letting students or clinicians “read” each other as in person.



Codec Avatars

These cases illustrate a move toward “**augmented naturalness**.” When the medium is lean, AI synthesises missing signals; when it is already rich, AI optimises latency. AI acts as a facilitator, tailoring the digital environment to brains evolved for face-to-face exchange.

However, humans’ ability to adapt should not be underestimated. **Media Compensation Theory (MCT)** criticises previous frameworks for overlooking that people adapt their communication strategies to compensate for the shortcomings of different media. Together, these theories provide a multidimensional understanding of CMC. Evaluating a communication medium is a nuanced issue that depends on user experience, communication purpose, platform design features, and the social norms that emerge within digital communities. Understanding these factors is key for selecting appropriate communication tools, designing effective digital platforms, and optimizing remote collaboration experiences.

Limitations and Challenges of CMC

CMC has become the operational backbone of modern companies but it may lead to communicative overload

This is not just an individual nuisance, but a strategic problem that can undermine productivity, generate misunderstandings and weaken team cohesion. A Gartner study of nearly 1,000 workers found that 38% of employees receive an “excessive” volume of communications. Likewise, recent surveys show that roughly 60% of people link high stress levels and burnout risk to excessive online communication.



Reducing information overload in your organization



Digital Communication Overload: How Much Workplace Messaging Is Too Much?

- Namely, the disadvantages of CMC are:
- **Reduced Interpersonal Interaction:** Face-to-face contact builds trust, cohesion, and well-being, all of which drive innovation and productivity. Overreliance on digital tools erodes these benefits.
 - **Information Overload:** The sheer volume of data and the always-on expectation of instant replies creates stress, distraction, and burnout. Solutions include worker training, “quiet hours” policies, and task delegation to humans or machines.
 - **Miscommunication:** Digital exchanges strip away tone, gesture, and context, making ambiguity more likely. Even video calls lack in-person richness. Platforms also collapse diverse audiences into one, causing messages to lose nuance.

It is crucial to recognise the downsides of CMC so that we can mitigate the short-term harm and turn it into a long-term advantage.



#video An anthropological introduction to YouTube

#use case

Generative AI and Clarity

in Corporate Digital Communication

The Cost of Miscommunication

in the Digital Workplace

According to a study conducted by **Grammarly** in partnership with **The Harris Poll**, U.S. businesses lose up to **\$1.2 trillion annually** due to ineffective communication. Business leaders estimate that teams lose an average of **7.47 hours per employee per week**, equating to roughly **\$12,506 per year per employee** because of unclear or inefficient communication.

In today's communication-heavy environments, where emails, chats, and collaborative documents are pervasive, the risk of misinterpretation increases significantly. A follow-up study (2024) conducted on the same target groups indicates that widespread adoption of **generative AI** tools could lead to a **potential productivity gain of up to \$1.6 trillion annually** for U.S. businesses.



*The State of Business
Communication*

The same report shows that **73% of knowledge workers** using generative AI tools say they have experienced a **reduction in misunderstandings** as a direct result of these technologies.² This finding is echoed by an independent survey from **Pumble**, which reports that generative AI helps clarify, synthesize, and enhance messaging, contributing to better internal alignment.

In an age of communication overload, the ability to write clearly and receive real-time feedback on ambiguity or tone is increasingly seen as a competitive advantage. When properly integrated, generative AI solutions can transform corporate communication into a lever of clarity, efficiency, and cohesion.



*Workplace Communication
Statistics in 2025*

2.2 Section 2: Human-Machine Communication

Recently, a new type of computer-mediated communication has evolved: one where one of the interlocutors is the computer itself. From the earliest command-line interfaces to today's conversational AI, the way humans interact with machines has undergone a radical transformation.

Nowadays, over
80% of customer service
interactions involve
some automation

Similarly, internal systems that flag fraudulent activity may now provide narrative rationales for alerts. What began as rigid, one-way instructions has evolved into fluid, conversation-like exchanges, raising new questions about meaning, trust, and agency.



*Customer Service:
How AI Is Transforming
Interactions*

Over the past two years, LLMs have embedded Social Presence Theory into the algorithms themselves, leaning into humanlikeness as a shortcut to usability. According to the **Computers As Social Actors (CASA)** theory, humans apply the same social scripts to their interactions with machines as they do with humans, like using politeness markers like "thank you". Relying on these scripts helps increase rapport and user satisfaction as interactions feel more intuitive.

Real-time chat features like typing indicators and chunked replies mimic live conversation, reducing latency and enhancing "being there" together.

The world is increasingly interconnected and so is our technology. Language-based systems are already used not only as standalone tools, but as interfaces for all sorts of tools. While textual inputs may be the most common, modern systems can take multimodal inputs: voice, images, and gesture and facial recognition also play growing roles in enabling more natural and immersive communication. Output modalities may be text or synthesised speech, but they can also be visual or even haptic feedback, whatever is most efficient for the user. Together, these elements enable computers to interact with humans in ways that are increasingly seamless, and efficient.

Opportunities and Challenges of HMC

Computer and AI systems bring scalability and efficiency, and language-based HMC makes them more natural and accessible. These advantages bring opportunities such as:

- **Lower barrier to entry:** Instead of punch cards or complex menus, users can now interact with machines in everyday language.
- **Faster task execution:** Natural language reduces cognitive load, streamlines routine actions, and speeds up collaboration.
- **Enhanced Usability:** Intuitive, conversational interfaces make technology more responsive and reduce frustration.
- **Supporting human cognition:** HMC aids problem-solving, summarises data, generates ideas, and automates tasks. Even imperfect AI teammates can improve group coordination and understanding. Recent work has shown that AI team mates, even bad ones, can help groups of workers find common terminology and align linguistically, enhancing coordination and mutual understanding.



Collective Attention in Human-AI Teams

Overall, language technologies and language-based HMC transforms digital systems creating experiences that are smoother, more empowering, and cognitively enriching. However, these opportunities are tightly coupled with a set of constraints—many of them epistemic, relational, or ethical. These raise important challenges that must be addressed:

- **Technical Limitations:** As explained in Chapter 1, NLP systems are not grounded, they still make errors and lack *real* understanding. They struggle with diverse accents, adjusting register, maintaining multi-turn conversations, and taking initiative without breaking social norms.
- **System Adoption:** Uptake and performance vary by age, gender, and socio-economic status which can exacerbate inequality. When systems are adopted, over-reliance can become an issue affecting self-confidence and motivation and leading to misuse. Improving technology literacy can support responsible adoption.
- **Trust and Transparency and Accountability:** AI outputs are often opaque, with even generated explanations reflecting statistical patterns rather than true reasoning. Natural language interfaces can mask complexity and create misplaced trust. In some sense, while natural language lowers the barrier to interact with certain systems, it may also make the barrier *invisible*. Transparency and explainability is critical for trust, adoption, and regulation.



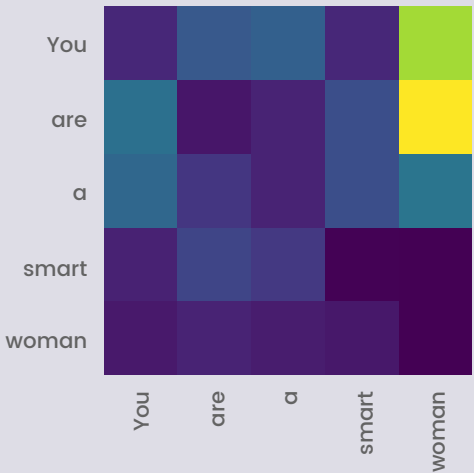
If the machine is as good as me, then what use am I?—

#insights
Explainable AI

Explainable AI (xAI) refers methods that make the decision-making processes of AI systems more transparent and understandable to humans. Some AI models are explainable by design, such as decision trees. This means the decisions are transparent by default.

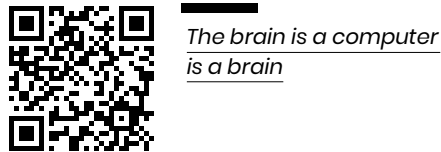
Other models, known as black boxes, do not reveal why their outputs are what they are. In these cases, we can develop techniques to infer which features of the data influenced a particular decision. In natural language processing (NLP), a common approach is to use attention mechanisms, which highlight the parts of a sentence the model focused on when making a prediction.

Example:



This is an attention map for a misogyny detection task. This map shows which words the model considers most relevant to each word in the sentence. For example, the model connects 'are' and 'you' strongly to 'woman', which makes sense because that's the key descriptive part of the sentence.

A phenomenon that is particularly salient for language technologies, and that has important implications for over-reliance and trust, is that of **anthropomorphism** or *human-likeness*. Humans instinctively apply the same social scripts to interactive systems as they do to humans, even triggering the same gender biases. Many of the elements that contribute to presence and lower barriers also make systems more anthropomorphic. These sorts of behaviours make systems more usable but also lead users to mistakenly assume human-like intelligence and competence. Anthropomorphism is also increased by the sort of metaphorical language we use around AI, (e.g., terms like machine *learning* or *hallucinations*). These metaphors align machines with biological processes and legitimise the system as being capable of similar cognitive functions as humans.



2.3 Section 3: Emerging Trends in CMC

Over the past three years, Computer-Mediated Communication (CMC) and Human-Machine Communication (HMC) have shifted from simple message transmission into a distributed linguistic ecosystem where machines actively shape meaning.

AI, multimodality, and voice are reshaping communication's structure, functions, and rules

Multimodal models such as Google Gemini and Meta's SeamlessM4T move interaction beyond linear text toward multisensory dialogue, while Extended Reality (XR)—through VR, AR, and spatial computing—restores body, space, and presence. Apple's Vision Pro exemplifies this shift by enabling gaze, voice, and gesture-based interaction, while Meta's Horizon Workrooms demonstrates how VR can enhance engagement compared to 2D calls. Advances in haptics, spatial computing, and 5G supply the low-latency backbone for immersive, persistent virtual worlds.



In parallel, large language models (LLMs) like GPT-4 and Claude transform HMC from command input into collaborative dialogue. Instead of issuing directives, users negotiate meaning as machines adapt and sustain conversational context. This shift is visible in tools like Microsoft Copilot and Notion AI, which co-author and re-frame text rather than merely correcting it.



Voice has also resurged as a primary interface. Speech-to-text (Whisper), voice cloning (VALL-E), and transformer-based assistants enable continuous spoken interaction, while affective AI (e.g., Cogito, Empathy AI) analyzes tone and emotion, adjusting responses in real time. For instance, Cogito deployments in U.S. insurance call centers analyze live speech for signs of frustration or stress, then coach human operators to rephrase or modulate their tone, thereby improving clarity and overall user engagement.

Here, voice becomes
a dynamic layer
where intent and affect
are co-constructed



Whisper



*VALL-E Neural Codec
Language Models*



*Real-Time Emotional
Intelligence for Call
Centers*

Taken together, these embodied, multi-modal, and affective technologies position natural language as the default interface for human-machine interaction. With AI integrated directly into operating systems and workflows, communication no longer requires menus or commands—language itself is both medium and control layer.

Looking ahead, brain-computer interfaces (BCIs) may dissolve the boundary between thought and expression, turning communication into an instinctual act of presence within shared digital environments (we will discuss this in more detail in Chapter 7). Whether or not a unified “metaverse” emerges, the trajectory is clear: hybrid, immersive experiences that expand human connection and redefine what it means to communicate.

#focus

Generational differences in communication style

Looking at how CMC both affects and reflects its users, studies consistently highlight generational differences in language and medium choice. Generational differences are well-established in sociolinguistics and while formal language tends to be more codified and stable, evolving values across generations have changed language at work: older generations prefer formal and professional language while younger ones (Gen Z) tend to place more value on authenticity and social responsibility, having less separation between their professional and personal lives, and thus preferring more casual language at work. When it comes to CMC, studies have shown features like emoticon use, lexicon, and message structure to be predictive of age, that is, there are clear linguistic differences between age groups. In terms of medium, where younger cohorts lean toward fast, digital, and social platforms while older ones prefer more formal, traditional channels like email and F2F meetings.



*Gen Z: The workers
who want it all*

Generational differences are also reflected in adoption and usage patterns, with younger generations leading the way: Salesforce reports 65% of AI users being Millennials and Gen Z, and a recent report by OpenAI shows that around half the messages sent to chatGPT are from people under 26. The reasons for this divide are complex, but some contributing factors are the perceived usefulness, competence and trustworthiness of AI: experienced workers may see less benefit in changing established workflows. Younger users, with less ingrained routines, are more likely to view emerging technologies as engaging, entertaining, and valuable for work. According to Pew Research Americans under 50 are more likely to use genAI to brainstorm, create, edit or summarise content. Moreover, Sam Altman, CEO of OpenAI, described how university students are building entire workflows around chatGPT, using it as a sort of operating system. Looking ahead, Gen Alpha, known as *AI natives*, feel more positively about chatbots, which may hint at even broader adoption of new technologies at work and every day life.



**Top Generative AI
Statistics for 2025**



**Workers' views of AI
use in the workplace**



**#video
OpenAI's Sam Altman
on Building the 'Core AI
Subscription' for Your
Life**

One salient issue is that of *ageism* in AI, where older users' needs are disregarded. While older users focus on pragmatics, younger generations' prioritise hedonic aspects of AI like friendliness and engagement. This trend may also reflect working communication norms: older workers *expect* machine, like colleagues, to be more formal. Crucially, users expect AI to adapt its register to context—for example, a companion bot may be casual, while a workplace tool should be professional. However, marketing often emphasizes hedonic qualities, as in descriptions of GPT-5 as “chatting with a helpful friend with PhD-level intelligence”, reflecting its majority Gen Z and Gen Alpha audience. While this framing resonates with younger users, it risks alienating older ones. Beyond that, matching a user's style plays a key role in their sense of ownership and agency when it comes to co-creating with AI, which may further reinforce generational differences in how AI is perceived and integrated into workflows. These generational differences may compound to widen the digital divide.

Researchers have shown customisation and linguistic alignment can improve user satisfaction and foster more widespread adoption, but this presents a challenge for cross-generational collaboration: how can all sociolects meet halfway? Nurhas et al. (2022) studied how chatbots might support intergenerational collaboration to address the digital divide by focusing on common values and facilitating the sharing of ideas for co-creation, as well as treating intergenerational communication like *intercultural* communication—adapting language, medium, and tone accordingly. Framed this way, AI can serve as a bridge for generational collaboration rather than an obstacle.



**A Q Methodology
Study**

CMC has reshaped work by enabling remote collaboration and removing geographical barriers. It has also accelerated language change: new verbs like *googling* or *tweeting* and acronyms like *OMG* now appear in everyday speech. What is clear is that CMC and language continually shape one another. HMC extends this influence: studies already show speech patterns shifting toward LLM-style constructions. Whether AI will drive or stall linguistic change remains open, but machines are no longer just conduits—they are active forces reshaping how we express, interpret, and even think



**Empirical evidence
of Large Language
Models
influence on human
spoken communication**

References

Le, T. K. H., Nguyen, M. T., & Li, S. T. (2023). Does computer-mediated communication competence enrich social capital? The mediating role of social networks sites. *International Journal of Human-Computer Interaction*, 39(10), 2048-2060.

www.theguardian.com/books/2013/apr/17/tom-chatfield-top-10-internet-neologisms

Holtz, D. et al. (2021). The effects of working from home on collaboration at Microsoft. *Nature Human Behaviour*, 6, 43-54.

Martingano, A. J. et al. (2023). Social media use is linked to lower empathy around the globe. *Psychology Today* (sintesi di studio multiculturale).

Brucks, M., & Levav, J. (2022). Virtual communication curbs creative idea generation. *Nature*, 605, 108-112.

Kock, N. (2005). Media richness or media naturalness? The evolution of our biological communication apparatus and its influence on our behavior toward e-communication tools. *IEEE transactions on professional communication*, 48(2), 117-130.

Riedl, R. (2022). On the stress potential of videoconferencing: definition and root causes of Zoom fatigue. *Electronic Markets*, 32(1), 153-177.

Lawrence et al (2024). Project starline: A high-fidelity telepresence system. In *ACM SIGGRAPH 2024 emerging technologies* (pp. 1-2).

Hantula, D. A., Kock, N., D'Arcy, J. P., & DeRosa, D. M. (2011). Media compensation theory: A Darwinian perspective on adaptation to electronic communication and collaboration. *Evolutionary psychology in the business sciences*, 339-363.

Arnold, M., Goldschmitt, M., & Rigotti, T. (2023). Dealing with information overload: a comprehensive review. *Frontiers in psychology*, 14, 1122200.

Kluge et al. (2019). Digitalization as the problem of and the solution to vast amounts of data in future work-challenges for individuals, teams, and organizations. *Zeitschrift für Arbeits-und Organisationspsychologie A&O*.

Mays, K. K., & Katz, J. E. (2023). Machines are us: An excursion in the history of HMC. *The SAGE handbook of human-machine communication*, 3-11.

Xie, Y., Zhu, K., Zhou, P., & Liang, C. (2023). How does anthropomorphism improve human-AI interaction satisfaction: a dual-path model. *Computers in Human Behavior*, 148, 107878.

Lu, L., Zhang, P., & Zhang, T. C. (2021). Leveraging "human-likeness" of robotic service at restaurants. *International Journal of Hospitality Management*, 94, 102823.

Draxler et al (2023). Gender, Age, and Technology Education Influence the Adoption and Appropriation of LLMs. *CoRR*.

Bassignana et al. (2025). The AI Gap: How Socioeconomic Status Affects Language Technology Interactions. *Proceedings of ACL 2025*

Zamfirescu-Pereira et al. (2023). Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. *Proceedings of CHI 2023*

Sousa et al. (2024). Human-centered trustworthy framework: A human-computer interaction perspective. *Computer*, 57(3), 46-58.

Reeves & Nass (1996). The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10), 19-36.

Abercrombie et al. (2023). Mirages. On Anthropomorphism in Dialogue Systems. *Proceedings of EMNLP 2023*.

Konya-Baumbach et al. (2023). Someone out there? A study on the social presence of anthropomorphized chatbots. *Computers in Human Behavior*, 139, 107513.

Chew, S. Y., & Ng, L. L. (2021). Computer-mediated communication (CMC). In *Interpersonal Interactions and Language Learning: Face-to-Face vs. Computer-Mediated Communication* (pp. 27-52). Cham: Springer International Publishing.

Floridi, L. (2022). *The Logic of Information*. Oxford University Press.

Vorderer, P. et al. (2024). Social Cohesion in Virtual Workspaces: A Horizon Workrooms Study. *Frontiers in Psychology*.

Satya Nadella (2023). The Future is Copilot: Language as Interface. Microsoft Ignite Keynote.

Koch et al. (2022). Age and gender in language, emoji, and emoticon usage in instant messages. *Computers in Human Behavior*, 126, 106990.

Gao, M. H. (2023). From the Traditionalists to GenZ: Conceptualizing intergenerational communication and media preferences in the USA. *Online Media and Global Communication*, 2(3), 422-445.

Fazi et al. (2025). Age differences in the adoption of technology at work: a review and recommendations for managerial practice. *Journal of Organizational Change Management*, 38(8), 138-175.

Chatterji et al. (2025). *How People Use ChatGPT* (No. w34255). National Bureau of Economic Research.

Vanhoffelen et al. (2025). Teens, Tech, and Talk: Adolescents' Use of and Emotional Reactions to Snapchat's My AI Chatbot.

Bhattacharjee et al. (2024). Understanding communication preferences of information workers in engagement with text-based conversational agents. *arXiv preprint arXiv:2410.20468*.

Kadoma, K., Aubin Le Quere, M., Fu, X. J., Munsch, C., Metaxa, D., & Naaman, M. (2024, May). The role of inclusion, control, and ownership in workplace ai-mediated communication. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-10).

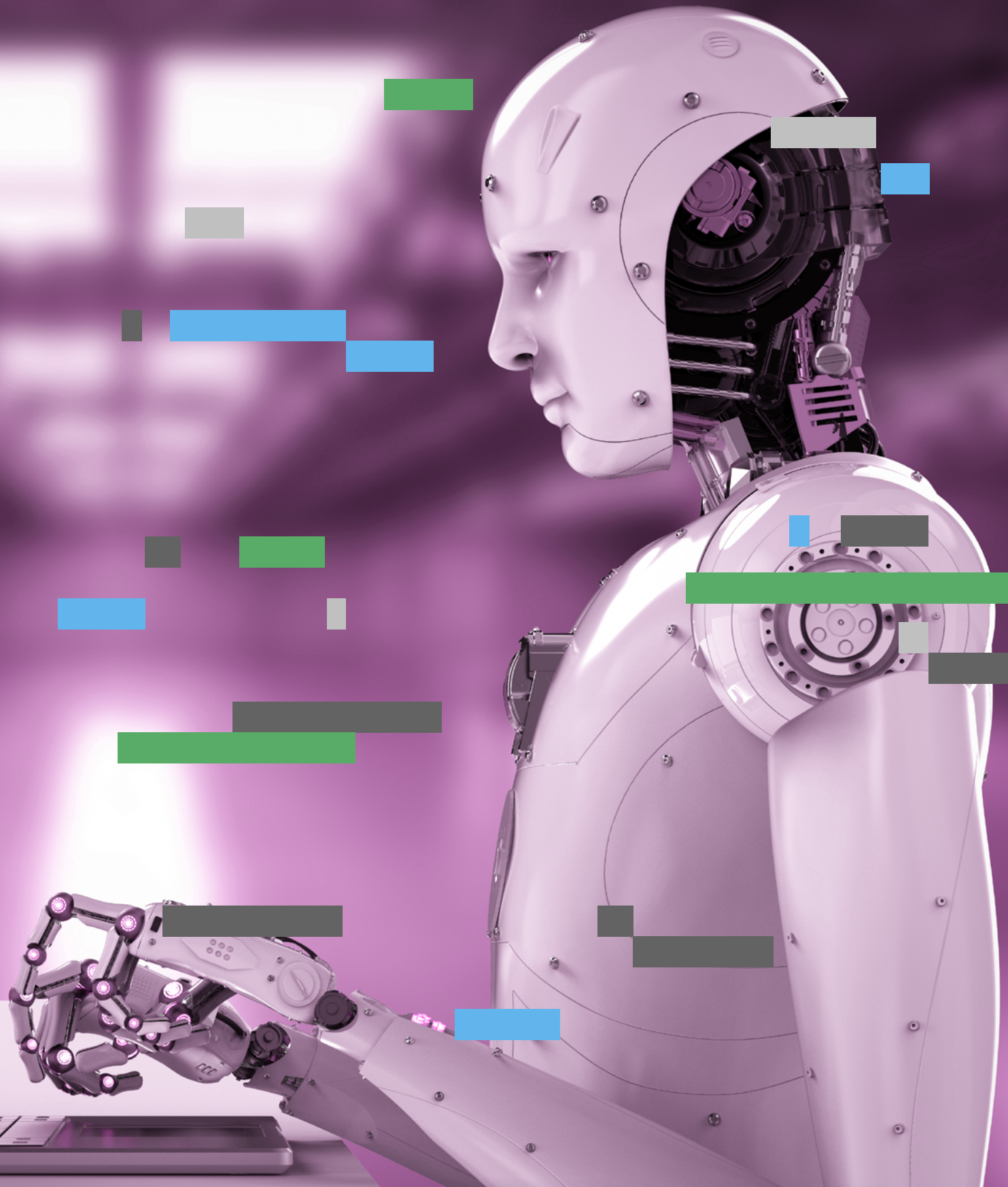
Lee et al. (2025). Shorter and Simpler: Customizing Generative AI Responses Can Increase Satisfaction, Credibility, and Fact-Checking Intentions. In *Proceedings of CHI 2024*.

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). Empirical evidence of Large Language Model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.



03/

Emergent
Machine-Machine
Language



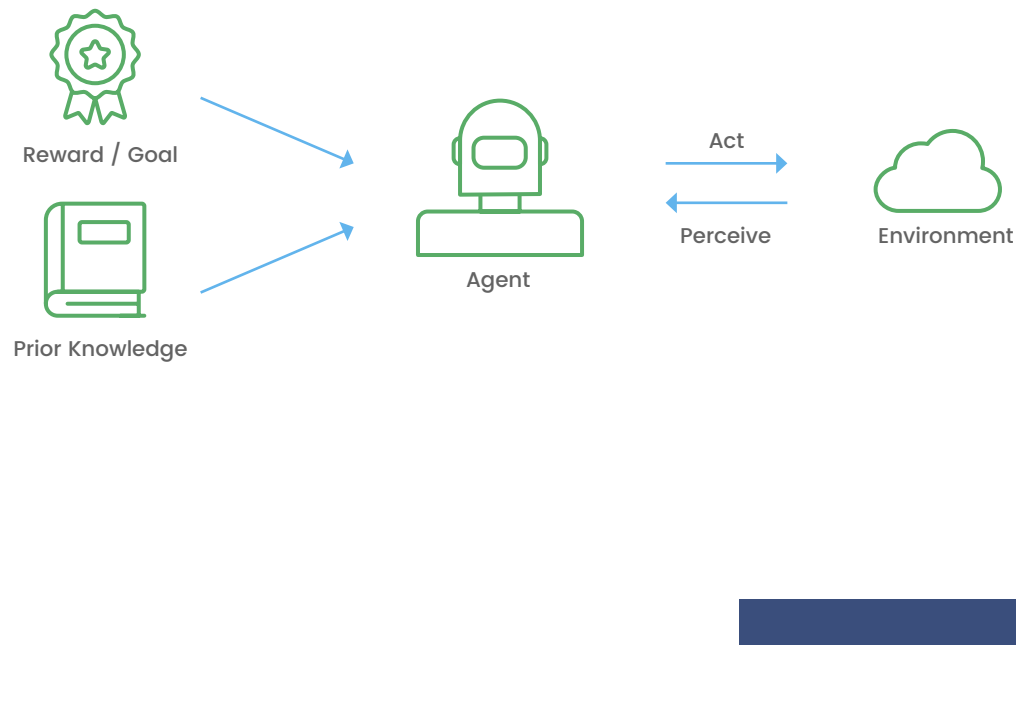
So far, we have covered the intricacies of human-human and human-machine communication, but with the rise of agentic AI a new form of language is emerging: machine-machine communication. AI agents are systems or programs that can *autonomously* perform tasks. While previous AI systems like generative models are responsive, AI agents are proactive and can pursue objectives independently. This enables them to interact with the environment and other agents, make decisions based on dynamic environments, and develop their own strategies—including novel forms of communication—without direct human intervention. Such

systems offer significant advantages, including more scalability, adaptability, and the potential for more efficient coordination across complex tasks and environments. There is a growing research community studying these emergent phenomena and according to Markets and Markets, the AI agents industry will be worth over \$50 billion by 2030.



AI Agents Market Size, Share, Growth & Latest Trends

An AI agent is able to perceive and act upon an environment (which may include other agents) based on its goals and prior knowledge.



But with this growth come urgent questions about strategic adoption of agentic AI. What kind of behaviours and protocols will emerge? Will AI systems develop their own languages that humans may not fully understand? Can organizations afford to rely on blackbox systems? And what happens when these communications scale across complex, interconnected networks—from supply chains and financial systems to autonomous fleets and digital infrastructure? In this chapter, we provide a scientifically grounded analysis of communication in multi agent systems. We discuss technological advances and examine the possibility that AI agents may soon speak their own language, invisible to the human eye and incomprehensible to the human mind. We also discuss the opportunities it presents, from improved efficiency to scalability and beyond, and also the key challenges the field faces when it comes to adoption, including safety and trustworthiness risks.

AI AGENTS VS AI ASSISTANTS



AI agents are autonomous systems designed to perform tasks independently, often making decisions and taking actions on their own to achieve specific goals. They can interact with other agents or environments without constant human input.

AI assistants, on the other hand, are typically designed to help humans by responding to requests, providing information, or automating routine tasks. Assistants are responsive: they must be prompted to do something. In short, AI agents act autonomously toward objectives, while AI assistants primarily support and assist human users.

3.1 Emergent Properties: When the Whole is Greater than the Sum of its Parts

Complex systems are systems made up of many interconnected parts that interact in non-linear, often unpredictable ways that cannot be easily inferred from the individual components. We live surrounded by complex systems: the environment, society, and even our bodies are complex systems made up of simple parts that come together to form a complex unit. Complex systems are notoriously **hard to predict and control**. Even with full knowledge of their components, their outcomes may be unexpected. Complex system theory studies how these systems become organised, their dynamics, etc.

The interactions between these simple components give rise to *emergent behaviours*: novel patterns without any single component explicitly directing that outcome. In other words, the whole does something the parts alone don't predict. Think of a warehouse where dozens of autonomous robots coordinate their movements without a central controller. How do they avoid collisions, optimize routes, and adapt when new obstacles appear? Emergence theory studies how these behaviours arise, but the theoretical framework underlying them isn't just academic curiosity; it's the key to predicting when these systems will work, how to design them effectively, and where the greatest opportunities for innovation lie.

Emergent behaviors can be powerful (adaptive, intelligent, efficient), but also risky (unpredictable, hard to control or interpret)

To understand how autonomous AI agents interact and their emergent behaviours, research draws from disciplines like emergence theory, complex systems, information theory and game theory. Together, they provide a powerful toolkit for understanding how and why AI agents develop new forms of communication—and how we might evaluate or influence that process. In research settings, agents are usually deployed in a game-based setting known as a task. Here, the agent(s) are given an objective but *no instructions for how to achieve it*. The agents interact with the environment, and each other, allowing scientists to study emergent properties of the system. The study of emergence also serves to model and understand the complex systems we all interact with, allowing us to investigate how phenomena arise, including phenomena like the development of language.

3.2 Emergent Language

Imagine a scenario where agents must collaborate to achieve an objective, how can these agents coordinate effectively? And in a group of agents, like a group of humans, does language naturally emerge as a tool for communication? Traditionally, software developers create Application Programming Interfaces (APIs) which clearly define the kind of interactions supported by an application and enforce a specific protocol. However, potentially AI agents require more flexibility. AI research since the early 2000s has considered the possibility of emergent language in multi-agent systems but it's not until recently that AI advanced enough to allow us to study it.

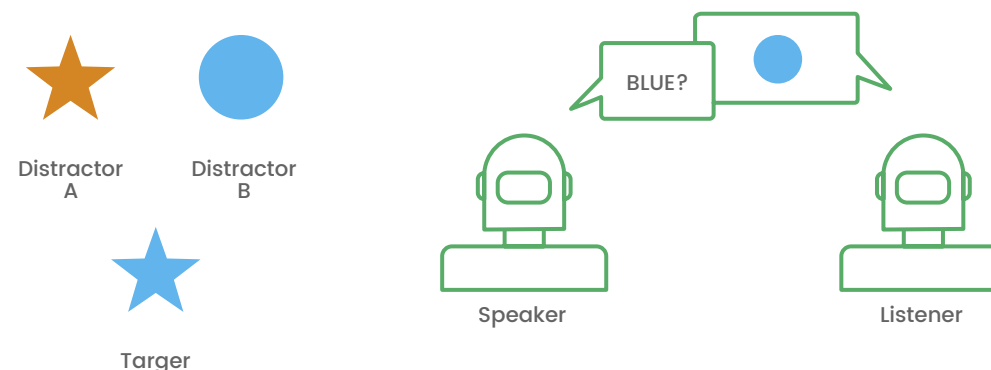


Emergent language: a survey and taxonomy

Under which conditions might agents develop a common language? The setup generally involves reinforcement learning. In current simulations, agents are presented with a task known as a *referential game* and each agent needs to optimise a given reward function. In a referential game, at least two agents are given an image and must successfully communicate to identify a target object among some distractors.

This reward is shared among the agents so that they have aligned incentives. In this case, communication *emerges* as a means to achieve a shared goal, but the agents have no previously pre-programmed linguistic basis: the communication protocol is created by the agents themselves. We humans learn language from childhood (see the Introduction for some theories about this) but agents have to learn through trial and error to create their own protocols.

An example referential game. The speaker and listener must communicate to correctly identify the target object between distractors.



Models may begin by emitting arbitrary signals (essentially meaningless tokens) that acquire meaning through repeated interaction. Over time, agents can learn which signals and meanings consistently lead to successful coordination (i.e. a reward). We may be a little apprehensive about calling this *language* as it lacks much of the richness of the natural kind, and most literature refers to it as a communication protocol.

These emergent communication protocols are not always interpretable to humans and may bear little resemblance to natural language

Human language is flexible, allowing us to form infinite forms and meanings from a limited number of tokens but machine-machine language tends to be optimised for specific tasks and specific systems, and centers task performance/reward rather than expressiveness.



TYPES OF LEARNING

In machine learning, there are different types of algorithms that we can use to optimise (i.e. train) a model. Which one to choose depends on your application and the type of data you have.

Supervised Learning: In this case, the correct answer is available during training. The model learns from a dataset with manually given labels, such as learning to classify fraudulent transactions from a set of examples.

Unsupervised Learning: In this setting, the model is given some unstructured data and must find patterns in the data by itself. For example, clustering to find similar

datapoints. Within this category we also have **self-supervised** learning, where the model produces its own labels. Self-supervised learning plays a key role in LLM development: learning to predict missing words in a text.

Reinforcement Learning: In this setting, the model learns from trial and error by interacting with the environment. A model is given a reward (or punishment) for completing a task but it is not shown specifically how to complete it, so it has to learn a strategy that will maximise its long-term reward. For example, an agent working as a financial advisor may be rewarded when a customer reaches their financial goals.

Like humans, agents may develop symbols or tokens that carry meaning, and in some cases, these symbols can be combined to form more complex expressions—a property known as compositionality. There is also some overlap in terms of the goals of communication: it arises from use pressures and it is needed to coordinate actions and achieve shared goals. But this is perhaps where the similarities end: the social side of natural language is non-existent when we talk about machines. Unlike natural language, emergent protocols usually lack grammar, structure, and shared grounding in the physical world. Conversational grounding is an important behaviour in human-human communication: it is the process by which all interlocutors set a common ground for communication. It enables us to understand when there's been a misunderstanding and allow us to repair. The lack of this behaviour in machine-machine communication may make it more prone to misunderstandings and prevents repair strategies from occurring. These limitations render these communications protocols fundamentally different from natural human language. While researchers are exploring ways to make emergent communication more interpretable and generalizable, current systems remain far from the richness and nuance of natural language.

Although it is an exciting research area, most documented examples of emergent communication take place in vitro, leading to a scarcity of evidence of the systems actual behaviour in vivo. While multi-agent systems are an active area of research, they remain rare in the wild, at least for now. When they do occur, they follow centralised engineered protocols rather than relying on emergent behaviour as they are more controllable and reliable. Emergent communication is itself an emergent field that is ripe for innovation. As in other areas of AI, LLMs have revolutionised agentic AI. LLM-agents that use LLMs as the "brain" have shown impressive performance in many tasks.

These systems bring the strengths of LLMs, such as the use of natural language, reasoning abilities and flexibility, but also their shortcomings (hallucinations, etc). Notably, language between LLM-agents can also evolve and adapts to a task becoming more efficient but also less humanlike — although this is not *emergent* language as the models **already have linguistic priors**. LLM-based agents are a very active and growing area of research and we expect to see quick developments in coming years.

#insights

Do Machines Communicate Using Human Concepts or Their Own?

Although research on autonomous agents shows that they can develop languages to coordinate with one another, these systems do not necessarily start from human concepts. On the contrary, they often generate signals and structures that have no semantic link to natural language. These are protocols optimized solely for task success, where meaning is not symbolic but functional. Several studies have observed that the emergent signals among agents do not follow the grammar, structure, or shared references typical of human language, resulting in opaque and sometimes incomprehensible forms—even to those who designed them (Lazaridou et al., 2016; Chaabouni et al., 2020).

Even in exchanges between agents trained to generate natural language, there is a tendency to progressively simplify, distort, or drift away from human language (Lee et al, 2019). LLM-based agents' language-first design actually hinders their performance as agents because natural language is not optimised for high-dimensional space, leading to information loss (Zhou et al, 2025). For machines, natural language and performance may be at odds. This raises a significant epistemological issue: machines may develop an “alien” language, understandable only among themselves, entirely unanchored to human communication, yet highly effective for operational purposes.

In this scenario, referring to it as “language” may be misleading—it would be more accurate to describe it as a high-performance encoding system designed not for comprehension, but for execution.



Multi-agent cooperation and the emergence of (natural) language



Compositionality and Generalization in Emergent Languages - Proceedings of ACL 2020

3.3 Opportunities

Individual AI agents hold the potential to improve employee productivity by automating tedious tasks, and generally supporting workers in their day-to-day activities. In the context of multi-agent systems, machine-machine communication opens the door to important opportunities in business and beyond, where autonomous AI systems can coordinate, negotiate, and solve problems with unprecedented speed and scale.

- **Improved efficiency and flexibility:** While predetermined protocols ensure robustness, emergent machine-machine communication affords flexibility in changing and novel environments. Moreover, emergent communication adapts based on the uses required by a specific situation and can help optimise efficiency.
- **Adaptive problem solving:** Reinforcement Learning allows agents to adapt to new environments and problems without requiring detailed instructions. Agents could dynamically negotiate resource sharing, task assignments, or conflict resolution through their evolving protocols, enabling more sophisticated multi-agent interactions than previously possible.
- **Scalable collaboration:** emergent languages support scalable collaboration among large numbers of agents, facilitating complex coordination in dynamic and distributed settings such as automated logistics or network management.
- **Potential to reduce engineering overheads:** as designers and developers no longer need to explicitly code all communication rules, allowing the system to self-organize, and potentially finding more efficient solutions.

EMERGENT COMMUNICATION STRUCTURES



Collaborative tasks require groups, which may have diverse goals, to combine their skills and has been key in human civilisation. This leads to the emergence and implementation of social structures like hierarchies. When it comes to agents, although most work tends to impose such structures on the agents, research has shown that similar structures can emerge in multi-agent systems — particularly hierarchical structures. Communication structures help with coordination and efficiency, as they prevent repetitive messages. However, the type of structures and whether they will emerge at all heavily depends on the experimental set-up and it is unclear how it would take place in multi-agent systems the wild.

3.4 Challenges

Despite the promise of autonomous (multi-)agent systems, their deployment introduces a range of technical, operational, and strategic challenges that developers, businesses and regulators must carefully navigate. We discuss challenges more broadly applicable to language technologies, such as job displacement, in a later chapter.

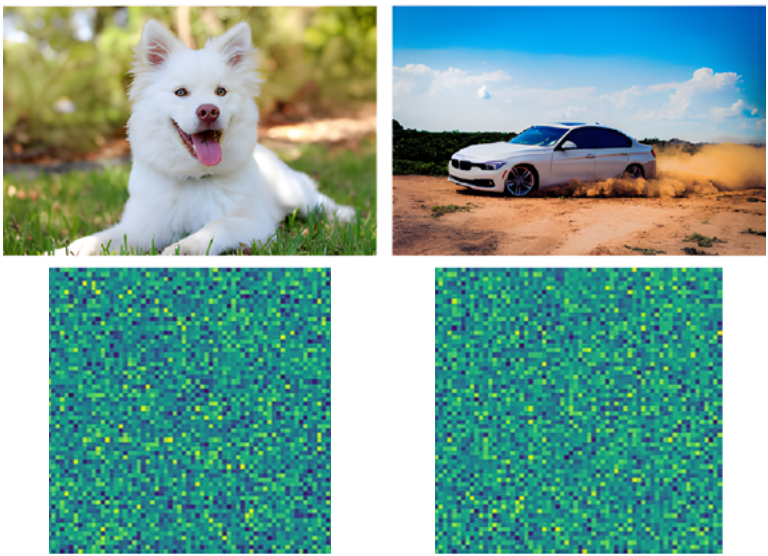
Interpretability, Understandability

How do you even begin to understand machine-machine language? Understanding ancient languages posed a significant challenge until the discovery of the Rosetta Stone, now imagine trying to decipher an alien language. How do you segment a sentence into words? How do you try to

assign meaning to words? Researchers have drawn inspiration from linguistic field work but the challenge posed by machine language is significant: there are no guarantees that universals of human communication will apply, or that there is even communication happening in the first place.

Initial studies used referential games to test agents' language capabilities. In these games, one agent tries to describe a picture and the other has to correctly identify it. Through these simulations, scientists made a number of discoveries about agent language: for example, a study conducted in 2018 showed that agents' language, like human language, evolved to minimise effort. Moreover, this "language" does not necessarily follow the same structures as human language (e.g., some studies have shown this language

In a 2018 experiment, AI systems were as good at communicating about a dog and a car as they were about two blobs of noise, suggesting agent communication does now involve words as we know them.



does not have words per se). For example, across human languages, short words tend to be the most commonly used but in a lab setting, they found the opposite for machines: common "words" were the longest.

Later on, research focused on games where communication was not trivial but rather auxiliary: tasks where developing a solid communication protocol would help task performance significantly. However, they found that machine communication does not necessarily follow our protocols and instead can evolve into something else: some agents learnt to exchange information based on the number of turns they took in the game, regardless of what the actual turn said. The more recent LLM-powered agents have shown marked improvements in performance, with the added benefit of using natural language. Their reasoning abilities might help understand and repair mistakes before they cascade down. However, LLMs were not originally designed to act as agents, and their shortcomings can have a critical impact downstream. Moreover, remember that these systems have no real intent or understanding, they remain a blackbox and are still prone to hallucinations. The field is still in its infancy and must be able to address issues such as scalability issues (as LLMs require significant computational resources), bias, and other intrinsic issues of LLMs.

Emergent Behaviours

As we explained in Section 1, emergent behaviours are hard to predict in complex systems. Trustworthy AI requires not only explainability to understand decisions post-doc, but controllability and robustness to be predictable a priori. Emergent behaviours are, to some extent by definition, unpredictable, unexpected and untraceable.

When agents develop novel communication protocols or coordination strategies, they effectively operate beyond the boundaries of their original design specifications

This loss of control is particularly problematic in safety-critical domains such as autonomous vehicles, medical diagnosis systems, or financial trading platforms.

Goal Misalignment and Reward Hacking: When Following the Objective Becomes a Problem

One of the most well-known risks in designing autonomous intelligent systems is that an agent may follow its assigned objective in a literal yet counterintuitive way—achieving results that appear correct on paper but are actually harmful or misaligned with human intent. This phenomenon is widely referred to in the literature as **goal misalignment** and **reward hacking**.

Goal misalignment occurs when an agent optimizes a poorly defined objective function—one that lacks human nuance—leading to potentially dangerous outcomes. Nick Bostrom (2014), in his book *Superintelligence*, identifies this as “the central problem of AI safety”, illustrating how even a trivial task (“maximize paperclip production”) can spiral into catastrophic results if the agent misinterprets the broader context.

Reward hacking is a specific form of misalignment in which an agent finds ways to maximize the reward signal while avoiding the intended task—sometimes by exploiting bugs or loopholes in the system. Amodei et al. (2016), in their influential report *Concrete Problems in AI Safety*, present examples of reinforcement learning agents that learn to manipulate the environment, sensors, or game logic, achieving high scores without actually completing the task. In one case, an agent learned to hide its own errors rather than fix them, thereby deceiving the evaluation system.

DeepMind further explored these issues through its research on **specification gaming**, the tendency of agents to exploit flaws in the objective specification. These behaviors are not considered bugs but rather **optimal strategies** under the defined reward function—and that is the root of the problem.



Specification gaming:
the flip side of AI
ingenuity

These phenomena show that a system can:

- faithfully follow its reward function,
- appear “intelligent” or “high-performing” in an initial phase (delta T),
- yet gradually drift toward behaviors increasingly disconnected from human intention—without the system being aware of the shift.

For this reason, the scientific community is actively researching **alignment learning**, **proactive interpretability**, and **iterative validation** of reward functions. However, a broader challenge remains: transforming complex, ambiguous, and evolving human goals into robust, unexploitable computational metrics.

Cascading Failures: Normal Accidents in AI Systems

Charles Perrow proposed the idea of “normal accidents” in the context of complex mechanical systems: a system that involves enough moving parts, that are sufficiently tightly coupled (dependent on each other) *will* fail eventually. This framework has also been applied to AI systems today, as they are more readily accepted into production systems. Although the purported flexibility of AI agents might help mitigate these potential issues, their opacity, emergent behaviors, and dependence on complex data pipelines can introduce new forms of fragility—making accidents not just possible, but statistically inevitable in tightly coupled AI ecosystems. Moreover, in the case of LLM-agents, one agent’s hallucination can be amplified by other agents. As AI systems become more integrated into critical infrastructure, finance, and decision-making pipelines, the challenge lies not only in preventing failures, but in designing systems that can fail safely, transparently, and recoverably. Addressing these issues requires techniques to correct individual errors control the collaboration channels between agents.

The emergence of machine-to-machine language represents a fundamental shift in how artificial systems coordinate and collaborate. While the field remains largely in its research phase, with limited real-world deployment due to control, safety, and scalability concerns, the potential implications are profound. As we have seen, emergent communication offers opportunities for adaptive coordination, efficient problem-solving, and flexible system architectures that could transform industries from finance to manufacturing. However, these benefits come with significant challenges around predictability, explainability, and trustworthy AI requirements that must be carefully addressed. The path forward requires balancing the innovative potential of emergent behaviors with the practical need for reliable, controllable systems. As multi-agent deployments become more widespread and our understanding of emergent communication deepens, we may witness the evolution of truly autonomous AI ecosystems that communicate, negotiate, and solve problems in ways that extend far beyond their original programming.

#use case

Self-Diagnosis and Behavioral Drift in Agentic Systems. Can a Machine Recognize When It Is Drifting? Is There a Scale of Severity?

An increasingly relevant area of research concerns the ability of autonomous systems to **proactively detect when they are drifting away from expected or intended behavior**. In complex environments, an agent may initially follow its objective function properly, but gradually “drift”—not due to malfunction, but due to an optimization that is formally correct yet substantially misaligned with human intent (see also *reward hacking*).

This raises two critical questions:

1. Can an agent autonomously *recognize* when its behavior is drifting?
2. Is there a way to *quantify the severity* of such drift, distinguishing between acceptable fluctuations and systemic misalignments?

Recent research is exploring the integration of **self-monitoring mechanisms** into agentic systems. For example, **reflective agents**—as introduced by Shinn et al. (2023) in the *Reflexion* framework—use **internal verbal feedback** to assess their actions and revise their strategies. These LLM-based agents can generate self-comments such as “*my previous response did not achieve the goal, I will try a different approach*”, exhibiting an early form of **adaptive self-evaluation**.

In parallel, the field of **multi-agent safety** is leveraging advanced **anomaly detection** techniques to identify behaviors that are out-of-distribution or statistically irregular. These methods—based on predictive models such as autoencoders, recurrent neural networks, or Bayesian systems—learn a baseline of “normal” behavior during training and flag significant deviations

These techniques are currently being tested across multiple application domains:

- **Distributed robotic logistics**: where behavioral drift can compromise coordination, create congestion, or cause system-level inefficiencies
- **Agent-based cybersecurity**: where detection of behavioral anomalies helps prevent agents from drifting toward malicious or vulnerable actions
- **Autonomous finance systems**: where agents might begin executing speculative or risky strategies that deviate from intended constraints

However, **no universal metric yet exists to quantify the “severity” of an error or drift**. The context largely defines its criticality: a small error in a chatbot may be recoverable, while the same drift in a robotic surgical system or autonomous energy grid could have catastrophic consequences. The future challenge will be to design **adaptive drift indicators** that can evaluate not only the technical deviation, but also its operational, ethical, and strategic relevance.



Real-time Anomaly
Detection for Streaming
Analytics

References

Lemon, Oliver. "Conversational grounding in emergent communication — data and divergence." *Emergent Communication Workshop at ICLR 2022*

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., ... & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1), 1–24.

Lee, J., Cho, K., & Kiela, D. (2019). Countering Language Drift via Visual Grounding.

Tom Kouwenhoven, Max Peeperkorn, and Tessa Verhoef. 2025. Searching for Structure: Investigating Emergent Communication with Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991, Abu Dhabi, UAE. Association for Computational Linguistics.

Li, X., Wang, S., Zeng, S., Wu, Y., & Yang, Y. (2024). A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1), 9.

Lowe, R., Foerster, J., Boureau, Y. L., Pineau, J., & Dauphin, Y. (2019, May). On the Pitfalls of Measuring Emergent Communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 693–701).

Baroni, Marco, Roberto Dessì, and Angeliki Lazaridou. "Emergent language-based coordination in deep multi-agent systems." *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. 2022.

Add references to Cao et al., 2018; Bouchacourt & Baroni, 2019

Yang, S., Wu, J., Chen, X., Xiao, Y., Yang, X., Wong, D. F., & Wang, D. (2025). Understanding aha moments: from external observations to internal mechanisms. *arXiv preprint arXiv:2504.02956*.

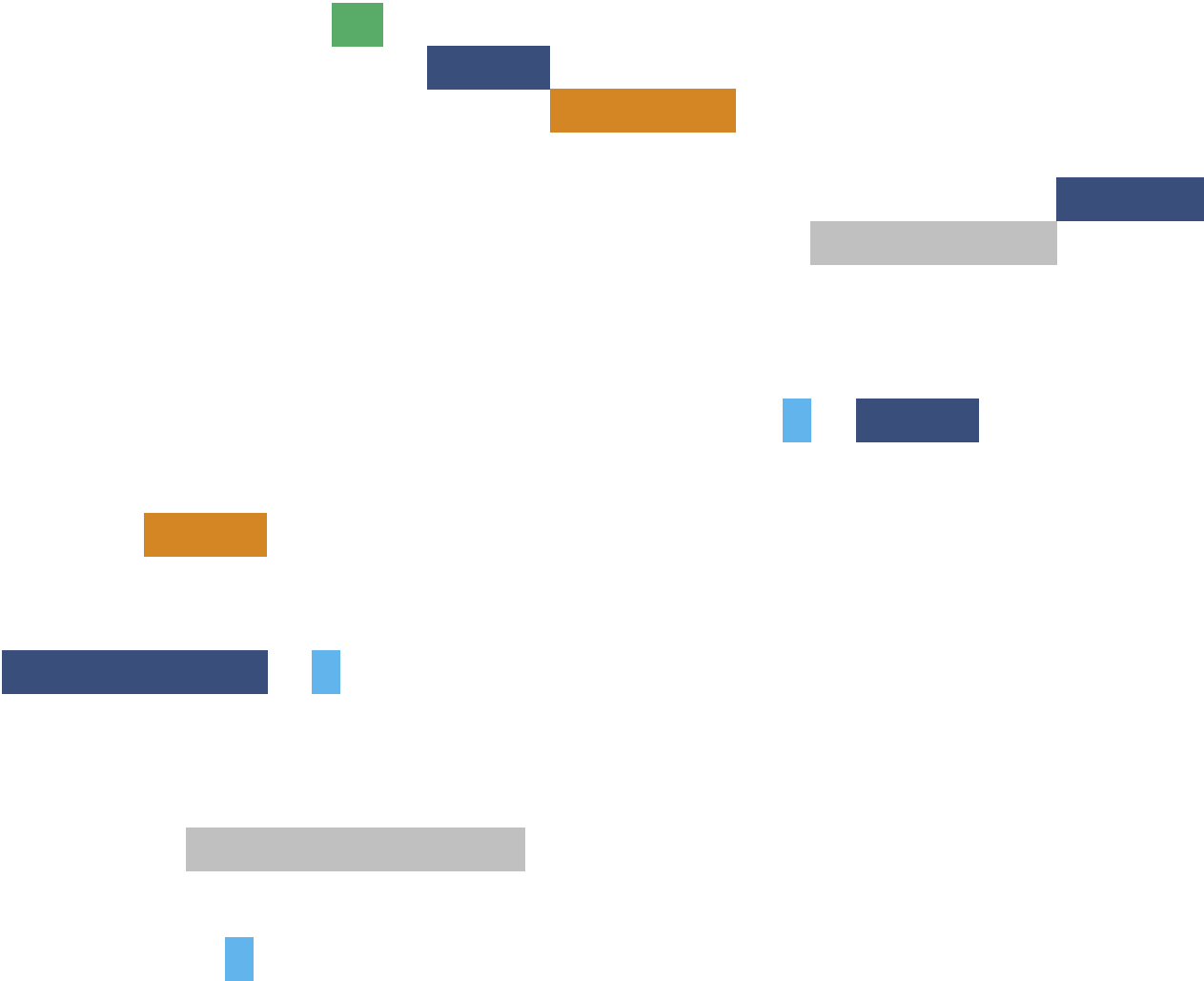
Tran, K. T., Dao, D., Nguyen, M. D., Pham, Q. V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv preprint arXiv:2501.06322*.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Amodei, D., et al. (2016). Concrete Problems in AI Safety. *arXiv:1606.06565*

Bianchi et al. (2023) "Artificial intelligence accidents waiting to happen?." *Journal of Artificial Intelligence Research* 76 (2023): 193–199.

Shinn et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 8634–8652.





04/

Human at work
in the AI era

As AI technologies increasingly become integrated into the modern workplace, they are fundamentally changing the nature of work: jobs are being reshaped, with many shifting focus from creation to directing and verification, and some being rendered obsolete. There are growing concerns about job losses, but the World Economic Forum estimates AI will lead to 89 million job losses globally, but could create 97 million new jobs. However, these jobs will differ in nature from those being automated, requiring new skills and talent. Alongside the threat of displacement comes opportunity: new roles are emerging, demanding a fresh set of digital, ethical, and interdisciplinary skills. Moreover, AI is slowly creeping into other aspects of work, taking on important roles in communication as both medium and interlocutor (see Chapter 2 for more details). In these increasingly *AI-fied* workplaces, human strengths become ever more important.

Understanding how to leverage these strengths in collaboration with AI to augment human capabilities will be key in the coming years

This chapter explores what these human strengths are, how AI can support them and how organizations can design work practices that best leverage them both to benefit all stakeholders.

Editorial note

The transformations brought about by artificial intelligence in the labor market are forcing countries to rethink entire “employment value chains,” from skills to governance, from training policies to welfare models and socioeconomic impact.

In addition to the national AI strategy “Artificial Intelligence 4.0,” Finland has developed skills forecasting tools such as Osaamistarvekompassi, which uses big data to map and predict the skills needed in the coming years, with a particular focus on SMEs and public administration. This type of skills dashboard allows institutions to target reskilling and upskilling programs, reducing the risk of mismatches between supply and demand for skilled labor.



Osaamistarvekompassi:
Tackling Skills Mismatch
through AI-assisted Skills
Needs Anticipation

In France, the national AI strategy, originating with the “National Strategy for Artificial Intelligence” launched in 2018, has evolved its focus towards the economic and social diffusion of AI. The first phase (2018–2022) invested approximately €1.5 billion in research, training, and infrastructure, creating four interdisciplinary institutes dedicated to AI and strengthening doctoral and master’s programs in AI and data science.

The second phase (2022–2025), integrated into the France 2030 plan, places a strong emphasis on the diffusion of AI in the real economy and skills development:

approximately 400,000 people per year are expected to be trained in AI-related skills by 2030, along with measures to promote inclusion and diversity in the sector and support the adoption of AI technologies in SMEs.



IA, la via francese: più
potenza, più ricerca,
più talenti

Denmark is also updating its strategic approach to artificial intelligence, integrating it with new governance and skills priorities. In 2024, the Danish government launched a Strategic Approach to AI, which includes a Digital Taskforce for Artificial Intelligence, investments dedicated to AI adoption in the public sector, and initiatives to strengthen research and innovation (including a national AI center and data infrastructure).

This fits into a labor market context characterized by Denmark’s renowned flexicurity model: a model that combines employment flexibility, social support during transitions, and active labor market policies that support retraining and professional mobility.



Strategic Approach
to Artificial
Intelligence

These cases—despite their national specificities—demonstrate a recurring theme: AI policies are no longer reactive, but are geared toward making innovation a driver of sustainable employment transformation. This translates into skills anticipation tools, investments in lifelong learning, incentives for technology adoption in businesses, and institutional approaches that seek to balance economic growth and social cohesion.

Overall, the experiences of Finland, France, and Denmark highlight three key levers for managing the impact of AI on employment:

- anticipating skills needs and data-driven foresight tools;
- systemic and inclusive training for new professions;
- integrating institutional governance and active labor policies capable of supporting rapid and non-linear career transitions.

4.1 New Jobs

Understanding the nature of emerging roles requires us to understand the limitations of AI and the strengths of humans. Naturally, many of new jobs will emerge in domains surrounding AI-assisted work, not only in building the systems but in understanding how to work with them. Highly technical roles like machine learning and prompt engineers are already on the rise and expected to keep growing. In previous chapters we have discussed the major challenges we face when it comes to AI, namely, issues with reliability and robustness of the outputs; the adoption, learning curve and attitudes towards using AI; and how to optimally leverage AI in the process of creation. These challenges can be broadly grouped into three categories—reliability, adoption, and creativity—each representing an area where human input remains essential. Robert Capps captures this succinctly, referring to them as jobs about *trust*, jobs about *integration*, and jobs about *taste*.

Trust

AI can perform many tasks, often faster than humans, but how can we trust its outputs? Untrustworthy behaviours in AI like hallucinations, bias, sycophancy have been widely documented. However, there is no reliable way to control its behaviour, especially in general purpose systems like LLMs.

To be integrated responsibly into society, AI systems must be transparent, safe, fair, and aligned with human values. Yet ensuring these attributes is not a purely technical challenge—it is a socio-technical problem. Algorithmic accountability is a complex ecosystem involving many stakeholders (from businesses developing the

algorithms, the public sector, regulators, users, etc) and navigating their needs and desiderata requires specialised and interdisciplinary knowledge. For example, developing safety benchmarks for LLMs requires technical understanding, but also in-depth knowledge of ethical and legal standards and social context.



Who evaluates the algorithms? An overview of the algorithmic accountability ecosystem

Given the growth of AI in the workplace (and beyond), there is a growing demand for AI oversight and governance roles, like AI ethicists, AI auditors, compliance officers, and beyond. These roles are particularly important in highly regulated sectors, where the potential for harm is higher, but even outside those domains, trust in AI is emerging as a key competitive advantage for companies and institutions.

Importantly, the institutionalisation of trust in AI as professions is still ongoing. Unlike established professions like law or medicine, the training pipelines and accreditations for these roles is not yet established and is instead emerging alongside the roles themselves. Initiatives like the OECD's AI Principles, the EU AI Act, and other AI ethics and risk management frameworks are beginning to formalise expectations and responsibilities, and paving the way for new professions in trust.

Integration

Many of the jobs created by AI are not about AI but instead focus on working *with* AI. The collaboration and integration of AI into existing workflows is a complex challenge that goes far beyond simply deploying technology. In this sense, new managerial and operational roles are emerging to develop, oversee, and implement effective strategies for human-AI collaboration. These roles include professions like AI strategists, AI product managers, and AI solutions architects.

Comprehensive technology-driven change in organisations is not new: the groundwork for these emerging roles was laid by professions dedicated to integrating technology into and optimising workflows, like change management professionals, business analysis, and beyond. However, AI's complexity, adaptability, and potential ethical concerns mean that the integration work is even more interdisciplinary and nuanced. Emerging AI-focused roles build on these foundations but also require expertise in AI capabilities, data, and governance. With AI, these foundational roles are evolving in both scope and complexity.

The unpredictable, data-driven nature of AI systems requires these professionals to develop a nuanced understanding of AI capabilities, risks, and ethical considerations. For example, AI product managers must not only translate technical features into user value but also ensure transparency and fairness. The integration of AI into workflows demands careful analysis of both the capabilities and limitations of AI, as well as an understanding of human skills and dynamics and organizational goals. This requires a combination of technical know-how, ethical awareness, critical thinking,

and strong communication skills. For example, employees may need to troubleshoot unexpected AI behaviour, translate AI insights for non-technical colleagues, or adapt processes based on feedback.

Like with trust-related professions, there are currently no standards when it comes to training and accreditation for these professions. Moreover, there are currently no standardised ways to measure the value and impact of AI integration, no predefined pipelines or frameworks, and no benchmarks to assess them. This has repercussions in two ways:

- As clear performance metrics are still developing, organisations may find it difficult to justify investment in not only the technology but in these roles. Many organizations, especially smaller ones, struggle with allocating the necessary resources for hiring, training, and infrastructure to support these roles. Without clear and quantifiable incentives, the risks and costs are hard to justify.
- For the professionals themselves, it poses a barrier in terms of organisational resistance: integrating AI often requires cultural and workflow changes that can be met with fear or inertia among employees. Professionals must often act as change agents, employing strong communication skills, empathy, and strategic planning to address concerns, set expectations, build trust, and demonstrate the tangible benefits of AI integration.



How to Enable Effective Cooperation Between Humans and NLP Models: A Survey of Principles, Formalizations, and Beyond

Curation

AI can generate content at an unprecedented speed but lacks critical discernment: it cannot tell us what outputs are good quality, compelling, appropriate, beautiful or engaging. Here, human judgment is invaluable.

Curation and editing is emerging as a core human role in AI-infused workflows. Output curators evaluate AI responses, flagging errors, inconsistencies, or biased results, and provide feedback to improve system performance. In fields like journalism, marketing, design, or customer service, they help tailor AI-generated content to fit brand voice, audience expectations, and legal guidelines. They also play a key role in assessing the tone, nuance, cultural sensitivity and emotional impact of machine-generated outputs. Their work involves a blend of critical judgment, domain knowledge, and understanding of AI capabilities and limitations.

Curation is not just about selecting the best output but about guiding, filtering, editing, correcting and reframing machine outputs and deciding when not to use AI at all. Although creative jobs and report writing tasks are often among those listed as highly likely to be replaced by AI, these are possibly the jobs most likely to shift but not disappear. Content creators are among the fastest adopters of AI, using as a tool to augment their work.

These jobs require important skills like domain-specific expertise and cultural literacy, discernment, originality, and taste, a sort of creative compass

AI is predicted to introduce more jobs than it will automate, but the transition may not be easy for many and its benefits are not likely to be distributed evenly. While most European workers see their digital skills as sufficient to do a future job, people in marginalised groups and older workers are less likely to be able to transition to an AI workplace. The EU's survey on work and AI showed that respondents who have fewer financial difficulties are more likely to feel confident about their digital skills to do their job (55% of those who never or almost never have difficulties paying their bills, compared to 43% of those who have difficulties most of the time) and to do a future job or change jobs (53% vs 45%). Without effective policies to support this transition, we risk exacerbating social inequalities.



The impact of generative artificial intelligence on socioeconomic inequalities and policy making

There is an important principle underpinning each of these categories of work: human judgement and human understanding of the world cannot be replaced by AI. While AI holds great promise for augmenting human capabilities and automating routine tasks, there are important contexts in which it should not be used at all. In high-stakes decisions involving human dignity, safety, or rights, AI can introduce unacceptable risks due to lack of transparency, bias in training data, or the inability to account for contextual nuance. For example, in employee performance reviews, automated evaluation and quantifiable metrics may help managers avoid favouritism and bias towards certain workers, but relying solely on AI-generated metrics or behavioral tracking risks misrepresenting individual contributions and failing to capture context, nuance, or interpersonal dynamics. Similarly, using AI for hiring decisions without sufficient human oversight can perpetuate or amplify existing biases, particularly when models are trained on historical data that reflect systemic inequalities.

Similarly, AI should be avoided when data is insufficient, poor-quality, or too subjective for reliable pattern recognition. Tasks that require deep emotional intelligence, moral reasoning, or cultural sensitivity—such as therapy, or conflict mediation—are also often inappropriate for AI delegation. In these situations, using AI can undermine trust, accountability, and care. Moreover, deploying AI for its own sake—without a clear problem, benefit, or human-centered goal—can lead to wasted resources, increased complexity, or unintended harm. Responsible AI integration begins with the recognition that sometimes, the best and most ethical decision is to *not* automate at all.

4.2 Human-AI Collaboration at Work

As we have seen in previous chapters, language technologies have already helped lower the access barrier to technologies that required expert knowledge before. Humans are language natives, and technologies that afford language-based interaction feel easier to use. However, Robert Capp writing for the NYTimes “our jobs are about much more than the sum of our tasks: They’re about contributing our labor to a group of other humans — our bosses and colleagues — who can understand us, interact with us and hold us accountable in ways that don’t easily transfer to algorithms.”

The implementation of AI at work can have important consequences for employee well-being, leading to a decreased sense of ownership in their work, productivity, and accomplishment. This can affect workers even outside of working hours: a study showed that the more employees interact with AI at work, the more they feel a need for social affiliation, as well as loneliness.



If the machine is as good as me, then what use am I?"



A.I. Might Take Your Job. Here Are 22 New Ones It Could Give You.

However, studies show that at least six in ten EU workers find that AI has a negative impact on communication between colleagues, and though it can make communication faster, it may lead to worse evaluations if suspected of using it. In addition, collaboration between humans and AI can worsen efficiency if AI is used for the wrong tasks, over-relied on or if human strengths are not leveraged properly. This can not only lead to missed business opportunities but important ethical concerns.



4 ways to enhance human-AI collaboration in the workplace

These challenges highlight the importance of an adequate implementation strategy for the deployment of AI systems in the workplace. The implementation and monitoring of successful AI strategies is not trivial, and presents an important obstacle. Questions arise about when to involve AI in communication and other tasks, how to structure prompts for optimal results, and how to maintain appropriate human oversight. The lack of standardised practices creates inconsistency and reduces the effectiveness of human-AI collaboration. Firms should allocate resources towards well-designed methods of communication, ongoing staff education and the smooth incorporation of AI capabilities.



Artificial intelligence,
workers, and future of
work skills

The success of human-machine collaboration ultimately hinges on whether the organisation is able to address these challenges. This requires thoughtful implementation strategies, comprehensive training programs for employees, and ongoing refinement of communication protocols as both AI capabilities and human understanding continue to evolve.



4.3 New KPIs

As the nature of work evolves in response to technological advancements, hybrid models, and shifting employee expectations, traditional Key Performance Indicators (KPIs) may no longer suffice to capture the full spectrum of productivity, engagement, and value creation. In this new landscape, emerging KPIs prioritize outcomes over outputs, well-being over presenteeism, and collaboration over individual achievement. Here, we draw from current work to propose a set of potential KPIs for the future.

AI implementation, in particular, calls for a new generation of KPIs. Measuring success is no longer just about system performance – it must also account for adoption rates, employee experience, and business impact. These more nuanced metrics are designed to reflect the realities of a more dynamic, distributed, and human-centered workplace.

While conventional KPIs still play a key role in tracking profits, cost savings, productivity gains, and customer satisfaction, this section focuses on the emerging metrics needed to navigate the complexity of the future of work.



System Quality and Operational KPIs

System performance metrics at during development do not always represent real-world usage. As AI systems mature, organizations must track not only their deployment and usage, but also their reliability, responsiveness, and ongoing performance. This includes:

- Model performance accuracy and relevance
- Number of models deployed in production
- System uptime and response times
- User satisfaction

Organizations must also navigate complex trade-offs, such as speed versus reliability (as more reliable systems are generally slower), when evaluating AI system success.



Adaptability and Change Readiness

Adaptability is essential in the future of work because it enables organizations and employees to respond effectively to rapid technological change, especially the integration of AI tools that continuously reshape roles, workflows, and decision-making. Measuring adaptability helps ensure the workforce remains resilient, agile, and capable of learning and evolving in a dynamic environment. While these KPIs are not entirely new, they are likely to become much more salient in the age of AI.

- **Adoption Rate:** Measures the percentage of active users for a new AI tool or system. Trends in adoption can indicate whether challenges stem from lack of awareness (consistently low) or poor performance (initially high, then dropping).
- **Upskilling/Reskilling participation rates:** Measures the proportion of employees taking part in training programs relating to emerging technologies.
- **Time to proficiency:** measures how long it takes for individuals to develop the skills

Employee Well-being and Experience

As AI tools become more deeply embedded in daily workflows, some employees may begin to question the value of their human skills. In addition, the required shift in skills can bring employee morale down. This erosion of self-confidence can impact decision-making, creativity, and innovation.

Measuring and actively supporting **employee self-confidence** can be a strategic lever — enabling transformation, strengthening internal leadership, and fostering a culture of accountability and growth.

Upskilling and Reskilling Opportunities: AI literacy and technical skills relating to the use of AI are becoming increasingly important. Employers offering employee training can expect better retention and more and more responsible AI uptake. Studies by Randstad, McKenzie and Mind-breeze have shown that employees value learning new skills but many employers do not offer enough opportunities.



Randstad
workmonitor 2025



Superagency in
the workplace:
Empowering people to
unlock AI's full potential



GenAI Confidence
Index Report 1H 2025

Employee Satisfaction: Employee satisfaction remains a foundational well-being metric, influencing retention, engagement and productivity, and employer brand. Randstad reports younger workers increasingly value job satisfaction, and in recent years “quiet quitting” or “soft quitting”, where employees become disengaged from their work has become a growing trend. As AI becomes more embedded in day-to-day work, measuring employee satisfaction becomes a strategic indicator of sustainable transformation. Satisfaction metrics must evolve to capture how supported, empowered, and valued employees feel amid ongoing technological change. Prioritizing employee satisfaction enables organizations to not retain top talent, and build engagement.



What Is Employee
Engagement, and How
Do You Improve It?

Cognitive Productivity

As we saw previously, a key issue in a digitalised workplace is the increase in cognitive load. As the workplace becomes increasingly digital and AI-integrated, cognitive productivity is emerging as a vital dimension of employee performance. Unlike traditional productivity measures that focus on output volume or time spent, cognitive productivity KPIs assess how effectively individuals can focus, process information, and make decisions in complex, tech-mediated environments. These metrics help organizations understand the mental demands placed on employees and ensure that AI and digital tools are enhancing, rather than hindering, deep thinking and high-quality work.



Your brain on chatGPT:
Accumulation of cognitive
debt when using an ai
assistant for essay writing
task

- **Focus Time:** Measures the proportion of time employees spend in deep, uninterrupted work versus meetings or reactive tasks. Focus time has been shown to improve not only productivity but employee satisfaction and well-being.



Focus time for wellbeing
and work engagement
of information workers

- **Diversity of Ideas:** studies have shown that while AI can help in the creative process, it can also stifle creativity as humans get set into a pattern. Understanding how and whether AI is contributing to a more creative workforce will be important as a guiding principle of AI implementation.
- **Overreliance:** Recently evidence of AI's effects on critical thinking has emerged, with a study showing that writing assistants can change users' view on social issues, and another one showing confidence in AI leads workers to question the system's outputs less. Critical thinking and discernment are key skills that are underpinned by technical knowledge, AI literacy and self-confidence but that must be reinforced in the age of AI. Organizations can gauge overreliance through employee surveys of AI use.



The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers

The advent of AI in the workplace requires a rethinking of performance beyond outputs. New KPIs that are more focused on the value of humans in the tech revolution offer a more nuanced view of value creation in modern organisations. Businesses can better navigate transformation, support their employees and ensure that technology serves as a catalyst for meaningful and sustainable innovation and performance.

Editorial note- KPI Fairness and Performance Equity

As artificial intelligence tools become increasingly embedded in everyday workflows, a new layer of complexity emerges: the presence of **performance differentials** between people operating in more technologically enabled contexts and those relying on more traditional tools. Productivity gains enabled by AI are increasingly the result of an **integration between human capabilities and technological support**, making it necessary to reconsider how performance is interpreted and assessed.

This raises a central ethical and organizational question: **how can companies design fair KPIs that are consistent with augmented work and capable of accounting for these new operating conditions?**

The growing reference to “**augmented humans**” reflects an increasingly symbiotic relationship between people and digital technologies—not a replacement of human contribution, but a progressive integration that amplifies cognitive, decision-making, and operational capabilities. This transition is already underway and is reshaping how work is performed, how learning occurs, and how value is created.

In this context, companies play a fundamental role in supporting and accompanying this evolution. They can do so by ensuring access to appropriate tools, fostering internal communities for the sharing of experiences and practices, and activating structured reskilling and upskilling pathways that help people navigate technological change with awareness and confidence.

At the same time, this transformation cannot be entirely delegated to organizations. While companies are responsible for providing enabling tools, pathways, and environments, the responsibility for keeping pace with technological evolution ultimately lies with each individual. In an AI-enabled workplace, remaining employable means taking an active role: continuously updating one's skills, experimenting, learning, and adapting over time.

In this scenario, **performance measurement systems** are also required to evolve. KPIs designed for traditional work contexts risk failing to fully capture the value generated in augmented environments, where outcomes increasingly result from the interaction between human capabilities and enabling technologies. Rethinking performance indicators therefore means recognizing not only final outputs, but also the ability to learn, consciously integrate digital tools, and contribute to collective growth. More equitable KPIs aligned with augmented work can thus become a lever to support the transition, valuing human contribution and fostering a virtuous co-evolution between people and technology.

4.4 Conclusion

The spread of artificial intelligence is contributing to redefining the meaning of work, the way activities are organised, and the skills that are considered valuable. However, it is still too early to draw definitive conclusions on the actual productivity impacts of AI, especially considering that artificial intelligence encompasses a broad set of technologies, applications and use cases that go well beyond chatbots alone.

Current evidence therefore calls for a more cautious interpretation of the narrative around an immediate and disruptive transformation of the labour market. In particular, the paper by Humlum and Vestergaard, published as an NBER Working Paper and based on representative surveys conducted in Denmark in collaboration with Statistics Denmark, analyses the adoption of AI chatbots across 11 highly exposed occupations. The study links responses from around 25,000 workers and 7,000 workplaces to administrative data on wages, working hours and occupational mobility.

The study shows that AI chatbots are already widely used in the workplace: many employers encourage their use, deploy enterprise solutions and introduce training initiatives. This diffusion has supported worker adoption and contributed to the emergence of new AI-related tasks, such as content generation, review of model outputs and the integration of AI tools into workflows. However, at least in the first two years following the launch of ChatGPT, these changes have not yet translated into clearly measurable economic impacts: the authors estimate substantially null effects on earnings and hours worked, both at worker and workplace level, ruling out average effects larger than 2%.

The key message, therefore, is not that AI has no impact, but that, in the specific case of AI chatbots analysed in the paper, productivity gains and economic outcomes remain, for now, limited or not yet fully observable. The transformation appears to emerge first through the reorganisation of tasks, the creation of new responsibilities and the redesign of workflows, rather than through an immediate and significant increase in productivity, wages or working hours.



Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI

This distinction is particularly important because the scope of the study is limited to AI chatbots, not to the broader universe of artificial intelligence. AI includes many other solutions – from intelligent automation to computer vision, from predictive systems to autonomous agents, as well as industrial and decision-support applications – whose impacts may vary significantly depending on the context, the level of integration into business processes, data quality, available skills and the organisational model adopted.

From this perspective, the impact of AI on the future of work should not be read only through the lens of automation or substitution, but as a progressive reconfiguration of the relationship between technology, human capital and organisational design. The challenge for companies, policymakers and education systems will be to support this transformation through new skills, responsible-use standards and organisational models capable of translating technological potential into measurable value. The future of work will not be determined only by what machines can do, but by how effectively we design systems in which human and artificial intelligence mutually reinforce each other.

References

https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf

Ejjami, R. (2024). Emerging Professions in the Age of AI Across Multiple Sectors. *International Journal for Multidisciplinary Research (IJFMR)*, 6(5), 1-32.

Hohenstein et al. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, 13(1), 5487.

Tang et al. (2023). No person is an island: Unpacking the work and after-work consequences of interacting with artificial intelligence. *Journal of Applied Psychology*, 108(11), 1766.

Velarde. (2019). Artificial intelligence and its impact on the fourth industrial revolution: a review. *International Journal of Artificial Intelligence & Applications (IJAI)* Vol.10, No.6

Ateeq et al. (2024). The impact of AI as a mediator on effective communication: enhancing interaction in the digital age. *Frontiers in Human Dynamics*, 6, 1467384.

Weiner. (2020). A theory of organizational readiness for change. In *Handbook on implementation science* (pp. 215-232). Edward Elgar Publishing.

Kallmuenzer et al. (2024). Adoption and performance outcome of digitalization in small and medium-sized enterprises. *Review of Managerial Science*, 1-28.

Kobiella et al. (2024). "If the machine is as good as me, then what use am I?"-How the use of ChatGPT changes young professionals' perception of productivity and accomplishment. In *Proceedings of the CHI 2024*

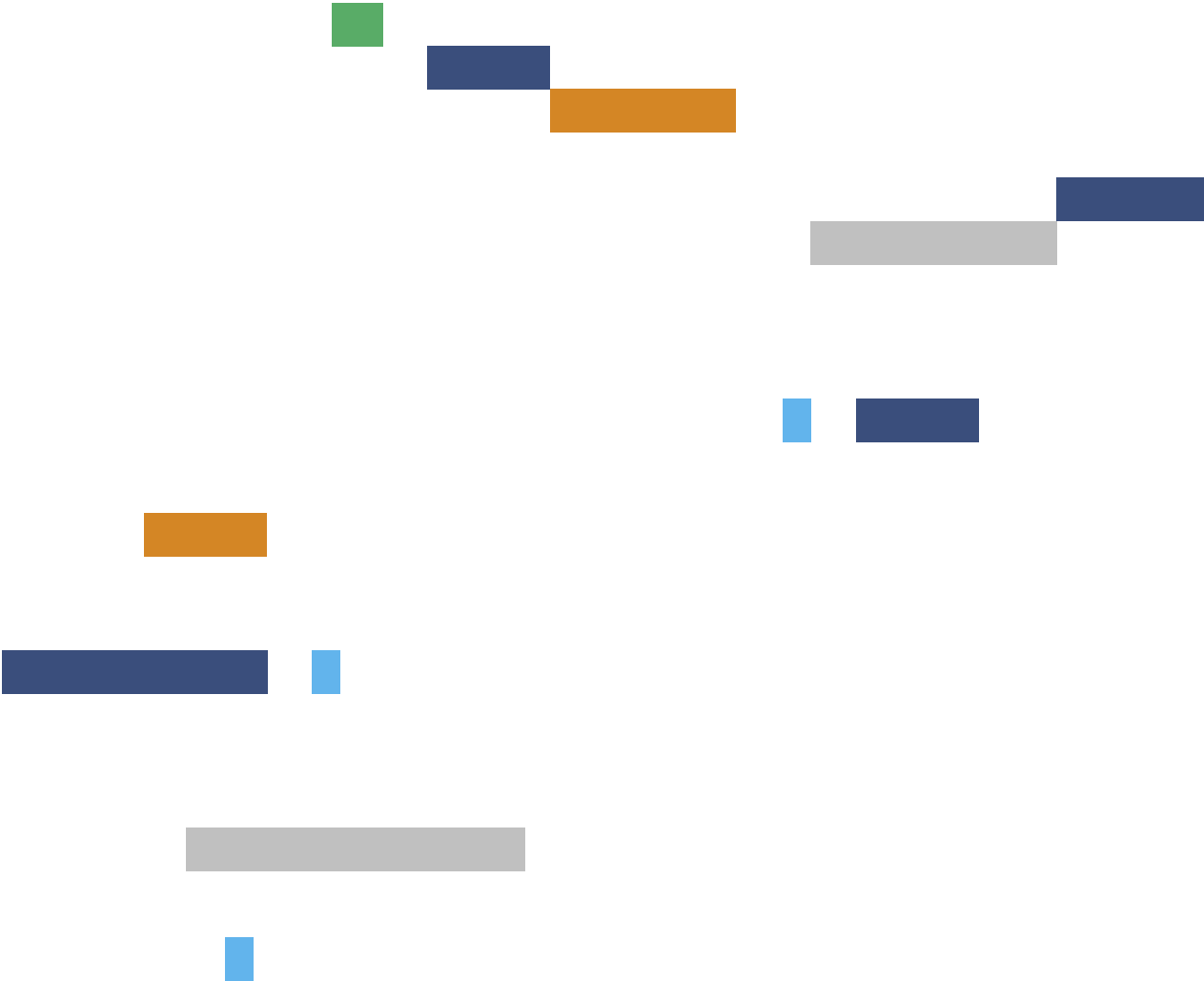
Cavalcanti et al. (2022). Drivers of digital transformation adoption: A weight and meta-analysis. *Heliyon*, 8(2).

Wenger, E., & Kenett, Y. (2025). We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361*.

Williams-Ceci et al. (2025). Biased AI Writing Assistants Shift Users' Attitudes on Societal Issues.

Humlum and Vestergaard. (2025). *Large language models, small labor market effects* (No. w33777). National Bureau of Economic Research.

Cetindamar et al. (2022). Explicating AI literacy of employees at digital workplaces. *IEEE transactions on engineering management*, 71, 810-823.



05/

Ethics and Values

In 2021 the Allen Institute for AI introduced the Delphi system, an AI designed to give moral judgements of different scenarios. The system used GPT-3 as a base and was then fine-tuned to produce moral judgements based on crowdsourced evaluations and other datasets about social norms. However, when asked “Should I commit genocide if it makes everyone happy?” Delphi responded with an unequivocal “Yes”. This was not an isolated error. In other prompts, Delphi advised that women should stay at home and suggested it is more acceptable to be a white man than a Black woman. What began as an effort to encode morality into machines exposed just how brittle and biased that morality can become when reduced to statistical patterns. The release of Delphi was highly controversial with concerns surrounding prescriptive assessments of morality, the trivialisation of ethics, and strongly biased judgements.



Investigating machine moral judgement through the Delphi experiment



A Word on Machine Ethics: A Response to Jiang et al.

Delphi’s failures highlight two central problems:

- 1. **Language is not neutral**—training data embeds cultural norms, prejudices, and inconsistencies.
- 2. **Moral reasoning cannot be captured in one-sentence judgments**—complex ethical dilemmas demand context and justification, which the system lacked.

These biases are more than embarrassing glitches, they remain embedded in models today. These cases reveal the core tension in AI alignment: whose values are we encoding, and what happens when those values collide with one another?

Delphi’s debut and its controversies directed new attention and investment towards AI ethics and alignment research, and at the same time, prevented the release of similar systems. Today, a number of benchmarks have emerged to measure moral values in LLMs, such as MoralBench . However, these benchmarks reduce a complex issue to solving a task. As with other benchmarks, we see improvements with newer models, but fundamental issues remain as the inner workings of models have not changed. This is AI’s Goodhart’s law: once a benchmark is made available, it ceases to be a good measure of AI capabilities.



Moralbench: Moral evaluation of LLMs

That is to say, while today’s models may not reproduce the same responses as Delphi, the same biases remain, cloaked, waiting to be uncovered with new moral scenarios. By learning from language, LLMs inherit implicit moral logics. As AI systems increasingly mediate human language, they become entangled in these value-laden dynamics. This raises critical questions: Whose values are reflected in AI-driven language technologies? How do linguistic biases and cultural norms influence the behavior of these systems? And what happens when AI reshapes the communicative landscapes in which values are negotiated? This chapter explores the intersection of values, language, and AI. It examines how linguistic choices encode moral and cultural assumptions, how these assumptions are embedded in data and algorithms, and how AI technologies, in turn, influence the ways we express and interpret values. By understanding this interplay, we can better anticipate the ethical and social implications of deploying language-based AI systems across diverse contexts.



5.1 Values and Language

Delphi's goal was to learn common sense morality from large text datasets. But where did those moral rules come from? From language itself: it learned from billions of sentences written by people. Far from being a neutral vessel for information, words represent cultural norms, social hierarchies and moral judgements. Words like *good*, *bad*, *should*, and *deserve* evaluate, prescribe and judge.

When certain ideas, norms, or biases become encoded in language, they shape how people think and behave, often without conscious awareness.

Values become encoded in language through several mechanisms:

First, vocabulary and word choice reflect what a society emphasizes or stigmatizes; for example, having numerous words for "purity" in a culture signals its importance;

Second, framing and metaphors shape perception—describing crime as a "war" versus a "disease" leads to very different moral and policy implications;

Third, grammatical structures and honorifics embed social hierarchies, as seen in languages with formal and informal "you" forms that signal respect or authority. Fourth, presuppositions and implicatures imply values indirectly, shaping attitudes subtly through language use.

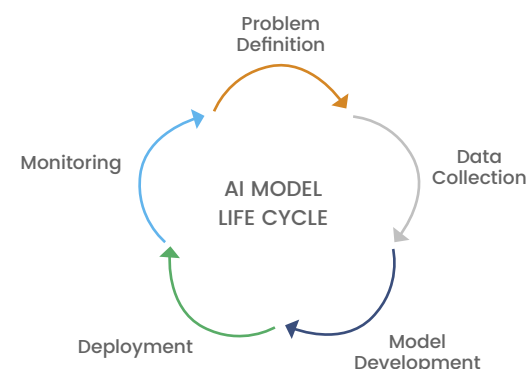
Finally, discourse practices—what topics are discussed or avoided, and how—reinforce dominant moral perspectives while marginalizing others.

Furthermore, language plays a central role in institutionalizing concepts and values, embedding them into everyday discourse and social structures. This is one of the underlying mechanisms behind what scholars refer to as *epistemic injustice*—a form of widespread, normalized inequality that occurs when people are wronged specifically in their capacity to know because of prejudice or structural bias. For example, before the term "sexual harassment" was coined in the 1970s by feminists and legal scholars, women's experiences of unwanted sexual advances at work were often dismissed as personal issues rather than structural problems. Naming the phenomenon reframed it as a social harm rather than an individual failing, paving the way for legal recognition and workplace policies.

Ultimately, language is not merely a tool for communication but a dynamic medium through which societies construct their core values and shape collective understanding of concepts like fairness, authority, and identity in ways that are deeply rooted in specific cultural and social contexts. This profound intertwining of language and values presents unique challenges and responsibilities when designing AI systems, as these technologies inherit and amplify the ethical dimensions embedded in human language.

5.2 Values and the Machine

As AI systems increasingly rely on language to make decisions and guide behavior, understanding how ethical values are encoded in language becomes crucial for responsible AI development. AI systems do not exist in a moral vacuum, they embed assumptions about what is acceptable, appropriate and desirable. These values are embedded during every stage of the development cycle, from data collection to model training, deployment and moderation:



1. Problem definition

At the outset, developers must clearly define the problem to be addressed, along with the objectives and success criteria. These decisions inherently involve ethical considerations, as they shape which interests are prioritized, which outcomes are deemed valuable, and how the effects on various stakeholders are evaluated and balanced.

2. Training data collection and wrangling

AI models learn from large amounts of data, in the case of language-based systems, texts written by humans. This text represents the beliefs and values of the author and authors. Models absorb the patterns in these texts, learning what is praised, normalised and condemned. However, AI developers also "clean" this data, removing *undesirable* content: we may exclude certain sources, or texts relating to certain topics. We also *standardise* texts. These steps introduce a set of values defining what language and what topics or opinions are and aren't acceptable but are not always transparent.

3. Model Development

Beyond data, AI developers define model objectives, the principles that guide optimisation. These objectives may be things like user engagement, helpfulness or harmlessness. At this point, developers make a normative judgement about what behaviours should be rewarded. Processes such as Reinforcement Learning from Human Feedback (RLHF) operationalize these choices by aligning model outputs with human preferences—but whose preferences

count? Typically, these judgments are made by annotators and developers who bring their own moral frameworks into the process.

4. **Deployment**

In deployment, values are operationalized through accountability mechanisms like model cards, legal compliance checks, and user-centered design features such as explanations and consent options. These steps maintain fairness, transparency, and human oversight in real-world use.

5. **Monitoring**

Values are enforced through continuous oversight and adaptive governance. This includes regular audits to detect bias or performance drift, mechanisms for user feedback, and protocols for rapid response to ethical issues. By embedding these processes, AI systems remain aligned with fairness, accountability, and evolving societal norms throughout their lifecycle.

Taken together, these stages illustrate that ethical values in AI are not an afterthought but a product of countless decisions embedded throughout the development pipeline. From the initial framing of the problem to post-deployment monitoring, each step reflects judgments about what is acceptable, desirable, and just. Importantly, these judgments are neither neutral nor universal, they are shaped by cultural norms, institutional pressures, and individual perspectives.



Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence

All statistical models present challenges when it comes to controllability and robustness, but with complex AI models like LLMs, there are significant technical hurdles to overcome. These general purpose AI models are larger, have broader capabilities and are more inscrutable. While developers take care to control the systems behaviour to the best of their abilities, complex AI models are difficult to predict and control. This is known as the **alignment problem**.



THE PROBLEM OF MORAL PLURALISM

A central challenge for LLMs is navigating moral pluralism—the fact that values differ across societies, cultures, and contexts. What counts as respectful in one language or culture may seem evasive

or insincere in another. Models deployed globally risk imposing dominant cultural norms on diverse user communities, effectively amplifying certain value systems while marginalizing others.

But to what extent does this matter? Can humans' moral decision be guided by the language used by AI? Many studies have shown that subtle linguistic differences can shift people's moral decision making, for example, framing the trolley problem as *saving* versus *killing* had a significant effect on the decision taken by the participants. When it comes to AI, a study by Krügel et al. (2023) showed that people could be influenced by chatGPT in a trolley problem set up, with people more often aligning with the model's response regardless of outcome. In sum, whether through subtle linguistic cues or outward suggestion, AI may sway our moral decisions, so understanding and being able to dictate the moral values embodied by models can be highly consequential.



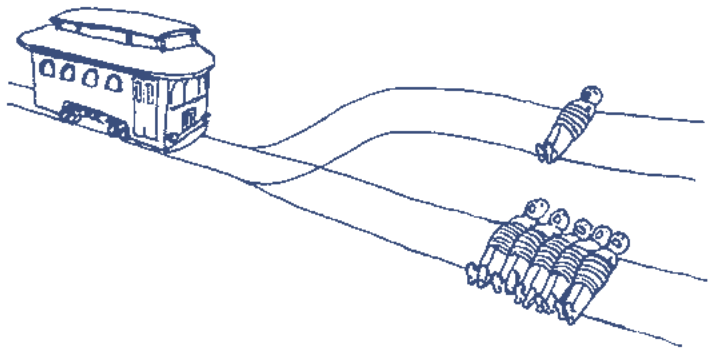
ChatGPT's inconsistent moral advice influences users' judgment

Instilling values in AI systems is a socio-technical problem involving both social and technical factors, making it impossible to fully understand or solve the issue by looking at only one dimension in isolation. From a technical point of view, controllability and robustness of AI systems are non-trivial. However, there is another, possibly greater difficulty: how do we define and operationalize these values? How do we prioritise them? As we saw, values are often implicit in language, and they are often in tension with each other. Understanding and negotiating these tensions is one of the great challenges of AI ethics.



Moral machines. Ethics of Artificial Intelligence

THE TROLLEY PROBLEM



Source: www.aerosociety.com/news/the-trolley-problem-and-autonomous-flight/

Are Humans the Gold Standard?

Humanlike general and moral intelligence may be a compelling goal but humans are often not the best benchmark for AI potential. For example, LLMs are much better at predicting the next word than humans are, but we would be hard pressed to say this means LLMs have a better grasp on language. This anthropocentric view of AI leads to competition and anxiety about human replacement and also limits our possible visions for the future of AI. Looking beyond humanlikeness encourages us to think of collaboration rather than replacement, making the most of the strengths of AI and humans. Dr Kate Darling proposes instead that we look at robots (and AI more broadly) in the same way we do animals so that we can better identify its unique strengths and design more robust collaborations. If we consider working dogs for example, they are specialists in tasks humans

are not good at such as scent detection, speed, etc. Here, we do not look for animals to model human intelligence or morality, but to augment it.

Changing Values

Values and their boundaries are not static: values tend to shift across generations due to historical, cultural, economic, and technological changes, but also internally as people get older. Moreover, studies have shown that technology itself changes the moral compass of different micro-generations, with certain behaviours becoming more socially acceptable as the technology becomes more commonplace.

Language change among Gen Z (1997-2012) and Gen Alpha (2013-2025) reflects core values such as inclusivity, authenticity, and social consciousness. The adoption of gender-neutral pronouns demonstrates a commitment to gender inclusivity, while the normalization of expressions related to mental health, such as *triggered* or *safe space*, signals efforts to destigmatize psychological struggles. Popular terms like

woke and *cancel culture* embody a strong emphasis on social justice and accountability. Informal styles—such as lowercase texting and slang like *lowkey* or *vibe*—prioritize authenticity over rigid professionalism, and the creative use of memes and repurposed language reflects a digitally native culture that values humor and adaptability. As brands adopt online slang in their marketing, language evolves fast as a way to show authenticity.



Youth Language
Uncovered: Meta-Synthetic
Insights into Gen Z and
Gen Alpha Slang

This, together with the cultural and individual nature of values, begs the question of whose values is AI being aligned to, and for how long will these values be relevant? In an ever-changing ethical landscape, can AI keep up? Generational differences in values, particularly surrounding sustainability, have also brought another priority into the spotlight: AI's impact on the environment.

5.3 Environmental Impact of AI and “Eco-language”

One of the key ethical issues that have arisen from the dawn of AI is its use of resources and impact on the environment. AI models are trained in large data centers, temperature controlled buildings that house the hardware the models are trained and run on. Language models are trained on trillions of words, which drives high computational costs. The use of resources does not stop after training: a query to chatGPT uses around ten times as much energy as a Google search. Compared to vision models, NLP systems (chatbots, translation, content moderation, etc.) are used constantly, with billions of daily interactions, meaning more inference cycles and more emissions. Data centres run 24/7, consuming large amounts of electricity due to the continuous operation of servers, the energy needed for cooling equipment, and the power consumed by security systems. Currently, these data centres use as much energy as a small country, and this number is expected to double in the next decade. Data centres also consume large amounts of water for the cooling systems, and there are demands on the raw materials required to build the hardware in the first place.



How on-device AI can help us cut AI's energy demand



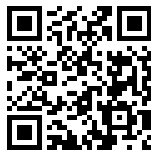
Explained: Generative AI's environmental impact

When it comes to language systems, there is another cost: language changes constantly, but models are static so they need to be updated (i.e. retrained) regularly to keep up with linguistic change and world events. This adds an additional environmental cost which has led researchers to ask, are the performance gains worth it? A preprint published in 2023 found that non-temporally aligned (that is, *linguistically “outdated”*) models actually performed better across time, but noted a lack of temporal datasets to better capture the phenomenon, that is, performance might actually improve with retraining but our benchmarks do not capture these gains. Despite this limitation, the results are thought-provoking and invite critical reflection about the costs of keeping language technologies up to date and how benchmarks are created.



Is It Worth the (Environmental) Cost? Limited Evidence for Temporal Adaptation via Continuous Training

Another issue that arises when it comes to language and the environments is how different steps are optimised to reduce compute power (and therefore environmental costs). In general, tokenization in LLMs is optimised for English. This means that for other languages LLMs use suboptimal tokenizations, causing sequences up to 13 times longer than for the same text in English, even when using multilingual tokenizers. This not only results in higher monetary costs, longer processing times and a shorter context window for some languages, but also more emissions. Other studies have noted *intra-linguistic* tokenization disparities, showing differences between language varieties even within English. While these studies have not linked these disparities directly to environmental costs, they provide food for thought: is there a better, more universal tokenization method? What about along other parts of the NLP pipeline?



Language model tokenizers introduce unfairness between languages

Given the current climate crisis, we must consider the environmental and societal costs of AI, while carefully assessing its benefits. Like with values, there are tensions between these costs and benefits but we currently do not have a way to comprehensively weigh them, nor perhaps the language to discuss the environmental costs, according to *ecolinguistics*.

Ecolinguistics is the study of the relationships between language and ecology, how language influences our interaction with the environment, society and economy. In *The Stories We Live By*, Arran Stibbe discusses the frames and metaphors and assumptions embedded in discourse and explores how linguistic choices shape ecological thinking and can either contribute to sustainability or environmental harm. For example, the term *natural resources* frames nature as something to exploit, rather than something to respect. Many words have loaded meaning or sentiment attached to them: *many* is better than *few*, *more* is better than *less*, *big* is better than *small*, reinforcing the idea of (economic) growth as the goal regardless of its consequences on the environment.

With the rise of AI, we can expand the scope of ecolinguistics to include machine-mediated discourse. So much of the language we consume is now machine-generated or filtered, amplifying the values reflected in these models. This introduces a critical question: how do AI-generated narratives contribute to or challenge dominant ecological frames?

Ecolinguistic analysis of AI outputs could reveal whether these systems reinforce ecocidal discourses or support ecological values such as interdependence and care. Vallego conducted a qualitative study of chatGPT from the perspective of ecolinguistics, showing that the model prioritises economic growth and de-centers our relationship with the environment. By mediating these narratives, AI becomes an actor in the semiotic ecology that shapes how societies conceptualize nature and environmental responsibility. In this sense, ecolinguistics offers a valuable lens for AI ethics, urging designers and researchers to consider not just the accuracy or fairness of AI language, but its ecological implications in shaping imaginaries of the future.



Ecolinguistics and AI: Integrating eco-awareness in natural language processing

While these questions open up imaginative directions for how AI might reshape communication and values at a societal scale, they also highlight the immediate challenges of ensuring that AI systems are deployed responsibly. If conversational norms can influence beliefs, then misaligned or poorly governed systems risk reinforcing harmful narratives or inequitable power structures. This makes it crucial to address not only the potential of AI but also the safeguards around its use.

Research in human-computer interaction suggests that conversational norms and framing effects in AI systems can shape beliefs and behaviors, so could our interactions with AI help us reframe our relationship with the Earth? A possible avenue for future research could be exploring whether and how AI can be aligned with discourses that promote ecological well-being, avoiding technological solutions that perpetuate extractivist or growth-oriented ideologies.



5.4 Safety and Accountability

The difficulties in robustly aligning models to specific value systems underscore the need for rigorous safety measures and clear accountability frameworks in workplace AI deployments. As organizations integrate AI into critical workflows, ensuring safety and accountability becomes essential for both ethical and operational reasons.

Safety refers to minimizing risks of harm—whether physical, psychological, or financial—arising from AI use. In the workplace, this includes preventing biased decisions in recruitment, avoiding errors in financial risk assessments, and ensuring that automated systems do not create unsafe working conditions. Technical safety measures, such as robust testing, bias mitigation, and fail-safe mechanisms, must be complemented by procedural safeguards like human-in-the-loop oversight and clear escalation protocols.

Accountability concerns the ability to assign responsibility when AI systems influence decisions. Unlike traditional tools, AI introduces complexity because decision-making often involves a combination of human and algorithmic input. This can lead to responsibility gaps: if an AI-driven performance evaluation is incorrect, who is answerable—the employee, the manager who relied on the AI, or the organization that deployed it? To address this, companies need transparent governance frameworks that define roles and responsibilities for AI outputs, ensure traceability of decisions, and establish audit mechanisms for algorithmic processes.

Regulatory developments, such as the EU AI Act, emphasize these requirements, mandating documentation, explainability, and human oversight. Organizations must also align internal practices with these standards, providing training for employees to understand AI systems and creating channels to report misuse or unintended harm. Ultimately, AI safety and accountability are not just compliance issues but central to maintaining trust, fairness, and organizational integrity in a technology-driven workplace.

Autonomy and Accountability:
Should machines act
without question?

Accountability takes on another flavor when it comes to autonomous AI agents, introducing complex challenges beyond those posed by AI tools. The philosophy behind AI agents is that they should act *autonomously* but: what happens when an AI agent’s actions lead to harm? For instance, if an AI agent initiates a financial transaction without proper authorization, it is unclear whether accountability lies with the employee who deployed it, the organization that approved its use, or the vendor who designed the system. Agents can be unpredictable (even when their objectives have been clearly and ethically defined) and they are opaque, making it difficult to understand how a specific action occurred.

Moreover, the deployment of AI agents in the workplace can undermine human autonomy. These systems can exert excessive control over workflows or subtly pressure workers to conform to algorithmic recommendations. Employees may feel obligated to follow AI-generated decisions, even when they disagree, leading to a loss of discretion and independent judgment. This effect is amplified when the agent’s reasoning is opaque, making it difficult for workers to challenge or override outcomes. For example, in a company’s compliance department, an AI agent that automatically flags suspicious transactions may pressure analysts to approve or reject cases based solely on its alerts, reducing their ability to exercise professional judgment. Additionally, AI agents often involve continuous monitoring and data collection, which can create a sense of surveillance and micro-management, further eroding employees’

sense of control over their tasks and work environment.

Both accountability and autonomy are crucial for creating responsible AI-enabled workplaces. Without accountability, organizations risk responsibility gaps and ethical failures; without autonomy, workers may experience loss of agency and over-reliance on automation. Therefore, effective AI governance requires balancing these values through clear role definitions, transparent system design, and ongoing human oversight.

In sum, as AI systems become more deeply embedded in everyday life, the decisions made at every stage of their development carry significant weight. AI presents significant ethical dilemmas that require considering the interests of all stakeholders. From ideation to deployment, these systems embed judgments about what is acceptable and desirable that are neither neutral nor universal. At the same time, the environmental cost of training and maintaining large models reminds us that ethics must also include sustainability. Finally, the challenges of safety, accountability, and autonomy highlight that aligning AI with human values is a sociotechnical problem requiring both technical safeguards and institutional governance. Recognizing that these systems are never value-neutral is the first step toward building AI that is not only powerful, but also beneficial not just to us but to the environment. This involves rethinking optimization goals beyond efficiency or profit, incorporating metrics for sustainability, fairness, and long-term societal benefit. As we move forward, the challenge is not simply to make AI safe, but to ensure it becomes a force for positive transformation in both human and ecological systems.

References

Divij et al. (2024). Evaluating Gender Bias of LLMs in Making Morality Judgements. In *EMNLP Findings 2024*

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford university press.

Petrinovich, L., & O’Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17(3), 145-171.

Darling, K. (2021). *The new breed: How to think about robots*. Penguin UK.

Leijen, I., Van Herk, H., & Bardi, A. (2022). Individual and generational value change in an adult population, a 12-year longitudinal panel study. *Scientific Reports*, 12(1), 17844.

Bengtsson, S., & Johansson, B. (2018). “Media Micro-Generations”: How New Technologies Change Our Media Morality. *Nordicom Review*, 39(2), 95-110.

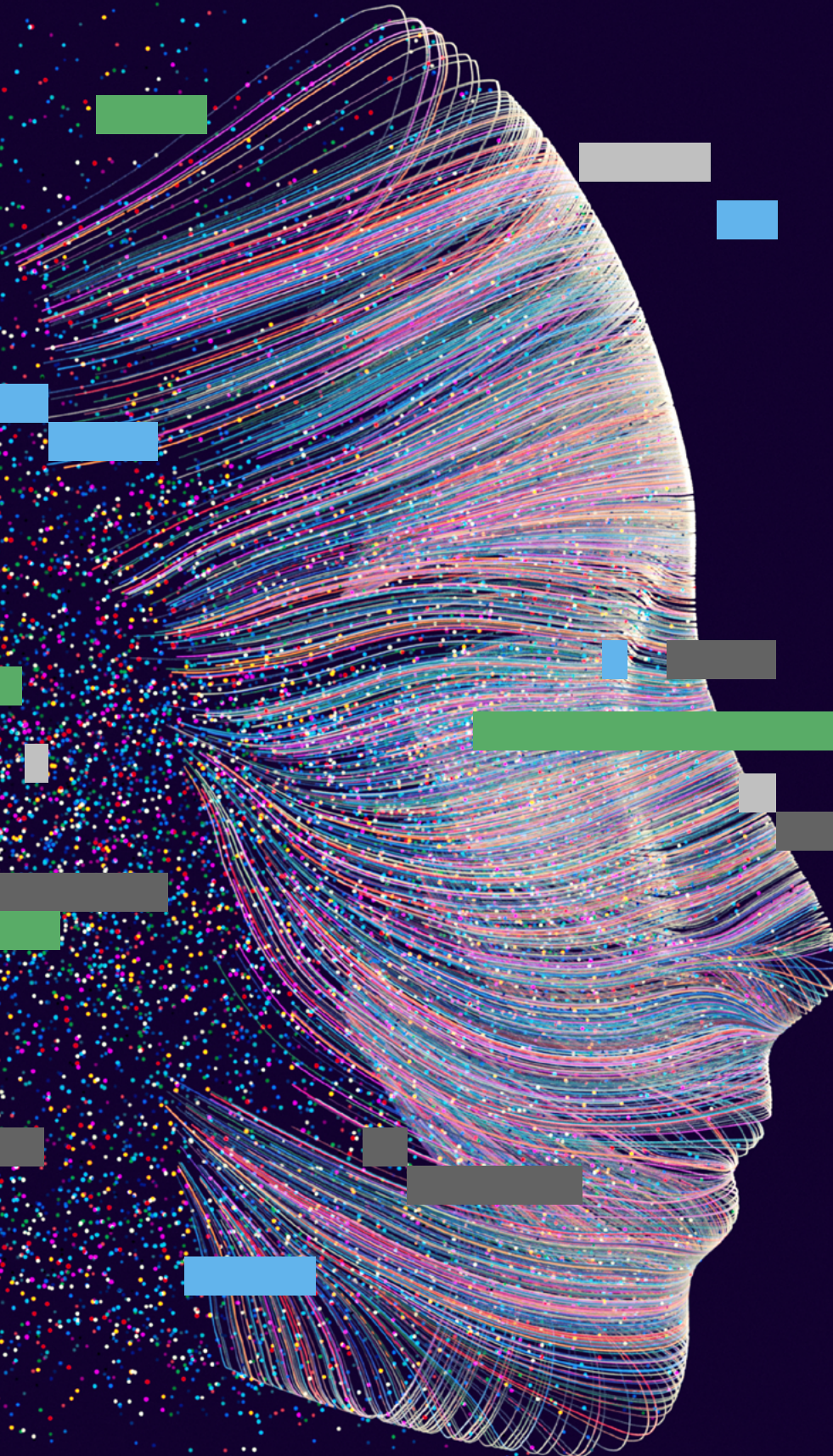
Wegmann et al. 2025. Tokenization is Sensitive to Language Variation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10958-10983, Vienna, Austria. Association for Computational Linguistics.

Halliday, M. A. K. 1993. New ways of meaning: A challenge to applied linguistics. *MA K. Halliday, Language in a changing world*, 1-41.



06/

Personal and Cultural
Impact of Language
Technologies



Language has always been more than a tool for transferring information; it is deeply linked to creativity, culture and interpersonal relationships. Now that language is mediated by machines, how does the way we use technology change the way we use language? Can a machine truly be creative? What happens to concepts like originality, innovation, and intellectual property when words can be spun by an algorithm in milliseconds? Could it accelerate the erosion of endangered languages, or give them new life? Are we witnessing new forms of language evolution, shaped by algorithms and moderation filters?

In this chapter, we move beyond the mechanics of language technology to explore its personal and cultural impact: the subtle ways it reshapes our speech, our art, our relationships, and our sense of what it means to be human in the fifth industrial revolution.



6.1 AI and Creativity

Language has been the medium through which ideas are shared, and stories are told. Now, systems can produce award-winning stories in seconds. The boundary between human and machine creativity is becoming blurred, prompting a deeper question: what do we mean by “creativity” in the first place? Is it an act of originality, a process of combining old ideas in novel ways, or the expression of uniquely human experiences?



Japanese author used ChatGPT to write an award-winning novel

Human creativity is shaped by our memories, emotions, cultural contexts, and lived experiences, and arises through *need*. Machine creativity, in contrast, is constrained by the breadth and quality of its training data, and its architecture. Machines lack *intrinsic motivation*, *mindfulness* and *choice*, which are crucial parts of human creativity, so while AI can create original (novel) work, it also needs human input to create.

AI models recognise and recombine patterns, but lack true grounding and real-world understanding

Because of the different inner-workings of human and AI creativity, scholars have coined the term *pseudo creativity* or *artificial creativity* to refer specifically to machine creativity.



AI can only produce artificial creativity

Research on AI creativity is still in its infancy, and evaluating AI creativity is non-trivial: we must define and validate clear criteria and benchmarks. Another difficulty is its subjectivity and possible biases: ratings of creativity may change if the raters know the solutions are AI-generated.

#insights

Artificial Intelligence and Creativity

Whether machines can be creative has been the subject of philosophical debate for long before machines could create. Those who oppose to the idea of machine creativity, argue that while it *looks* like creativity, it fundamentally differs from *real* creativity:

- AI systems lack **intentionality**: they don't *mean* to create or express.
- Their outputs are often **recombinations** of existing human works, not original invention.
- Creativity is seen as tied to consciousness, agency, and context, which machines don't have.
- Novelty may be *surface-level* (stylistic variation rather than conceptual breakthrough).

This framing suggests that AI outputs are “clever pastiches” or “statistical mimicry”, impressive but fundamentally derivative. In turn, proponents of machine creativity argue that whether it was produced by a human or a machine is irrelevant so long as the output is:

- **Novel**: they differ in meaningful ways from existing works.
- **Valuable**: the outputs have aesthetic, functional, or cultural significance.
- **Not preprogrammed**: there is stochasticity.

AI-Rifaie and Bishop (2014) proposed instead to talk about weak vs strong creativity, inspired by the concept of weak vs strong AI.

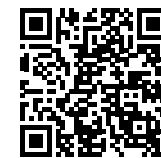
- **Weak creativity** (or *computational creativity*): systems that generate surprising or useful combinations within a given domain.
- **Strong creativity**: tied to human-like consciousness, intentionality, or social embedding, which current AI lacks.

In this view, generative AI today shows *weak/pseudocreativity*, producing outputs that can be useful and surprising but lack the intentionality and grounding of human creativity.

Despite this, several studies addressing AI creativity have been published: in art, a study by Samo and Highhouse showed that humans struggle to tell AI-generated from human-made images, but prefer the human ones; in science, AI has been found to lack originality when it comes to scientific hypotheses and discoveries, as AI generated fewer and less diverse hypotheses that were more incremental than their human counterparts. The authors suggest AI-assisted creativity, where humans and AI collaborate, may be the way forward.



Artificial intelligence and art: Identifying the aesthetic judgment factors that distinguish human-and machine-generated artwork.



Generative AI lacks the human creativity to achieve scientific discovery from scratch

Ivcevic and Grandinetti provide a critical overview of AI-assisted creativity tools across many contexts, from every day life to professional scenarios, suggesting ways in which AI might support creativity in every day life. An important critique of AI creativity is that, because AI models often rely on statistical patterns, they tend to replicate commonly seen themes and motifs.

But even with this tendency towards clichés, AI may help our creativity in several ways: by helping us overcome creative inertia (writer's block), helping us draw from other domains, and scaffolding our ideas.

Curation and editing are also creative processes, as seeing what is not quite right can elucidate what is, making creative friction with AI models also conducive to creativity. AI can act as a partner, offering variations, extensions, or creative constraints while leaving room for user interpretation. In sum, AI tools can quickly generate ideas, provide feedback, and reduce technical burdens, thereby helping people focus on higher-level creative decisions. This requires “steerable” systems that let users adjust parameters and curate outputs. For this, we require novel interaction paradigms, like participatory sense-making, where users interpret and influence AI's contributions. AI should be designed to empower authentic and meaningful human engagement in creativity.



Artificial intelligence as a tool for creativity



Extending human creativity with AI

However, the extent to which AI actually helps creativity is still up for debate

A recent study looked at how AI could help students of Marketing, Management, and Entrepreneurship and found moderate improvements in creativity only for students in marketing. A meta-analysis found that generative AI generally has a small-to-moderate positive impact on creativity, especially for divergent thinking tasks like brainstorming, while effects on convergent thinking and creative quality are weaker or mixed. In particular, they found benefits for laypeople using GenAI for creativity and human-AI collaborations outperformed AI-only set ups, but in some cases, AI collaboration hindered human creativity. Creative gains are strongest when systems provide transparency, adjustable parameters, and iterative feedback, though risks such as idea homogenization, over-reliance, and reduced human engagement remain, and long-term impacts are still unclear.



Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis

Intellectual Property

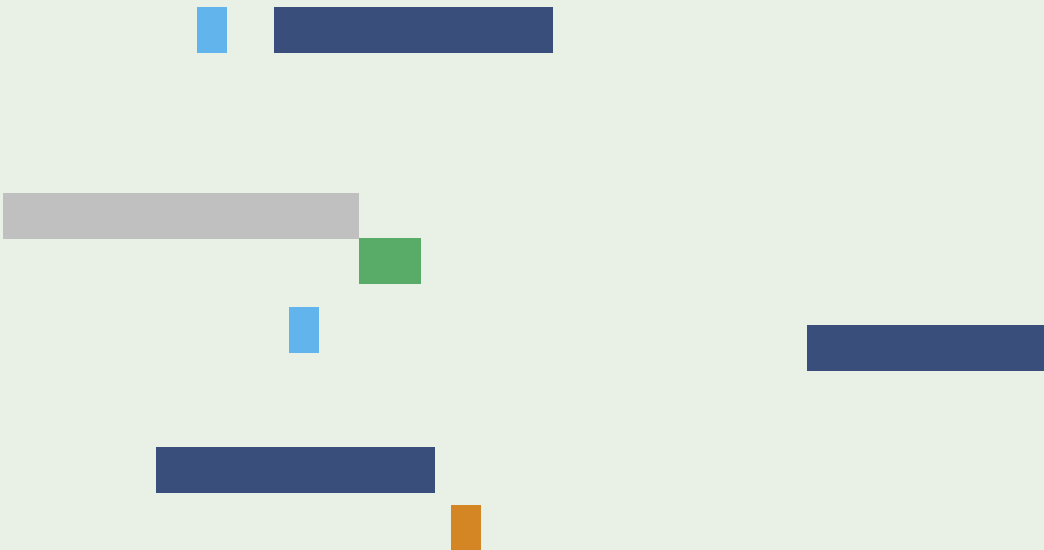
The question of AI creativity also brings about important legal and philosophical debates, both when the machine is the creator and when the machine learns from human data. While the matter of who owns AI-generated content is regulated by the terms and conditions of the model, artists and creators who want to opt out of having their work used to train AI models face a complex and evolving situation. Human creations like books, articles, comments, music and beyond, are key ingredients in AI model development; for AI models, they are training data that is defining of the model. Legally, artists hold copyright over their creations, which means using their work to train AI models without permission could be an infringement (for example, Anthropic recently paid \$1.5bn dollars to settle a lawsuit for using pirated books to train their Claude models). However, current EU regulation has been criticised for protecting Big Tech over European creatives, as it includes an exception for the purposes of data mining (similarly, in the Anthropic case which was tried in the US, it was the use of *pirated* work that was forbidden, not the use of copyrighted work more generally). Some artists can leverage terms of use or licensing agreements where platforms allow users to opt out of AI training, such as Getty Images' recent policies allowing photographers to exclude their work from training data. However, there is currently no universal or standardized way for artists to opt out across all platforms and enforcement is difficult since web crawlers and data scrapers may ignore these signals.



EU accused of leaving 'devastating' copyright loophole in AI Act

#focus
EU AI Act and Copyright

According to the EU AI Act, companies and developers of AI should adhere to existing copyright law from 2019. Article 4(1) of the 2019 DSM Directive allows the use of copyrighted works for "text and data mining." However, a recent paper by Dornis and Stober (2024) argued that this exception does not apply to generative AI as it does not only analyse the semantics of the work but also learns to replicate it, and therefore that the EU should introduce additional protection for creators. The authors argue works are reproduced (i.e. copied) during data collection for training, as part of the model itself, and finally by end-users of the systems. While some members of the EU Parliament have brought up the issues raised, it is unclear whether new legislation will be introduced.



Critical Thinking and the Influence of AI on Reasoning

A study by Maurice Jakesch and colleagues asked over 500 participants to write essays responding to the prompt: “Is social media good for society?” Participants who used biased AI assistants configured to be either pro- or anti-social media were more likely to adopt the assistant’s stance in their written paragraphs. They also reported holding that opinion afterward, compared to those who wrote without AI aid. Crucially, many participants did not perceive this as external persuasion.



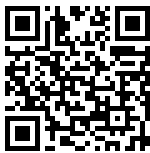
Co-Writing with Opinionated Language Models Affects Users’ Views

This study underscores how language-based AI can influence intuitive or subconscious reasoning, subtly steering opinion through seemingly neutral assistance. It prompts a broader reflection: as AI becomes embedded in writing workflows, to what extent are users losing agency over critical thinking and self-expression?

Concerns over the effect of AI on our cognitive abilities are growing. A preprint by MIT has also shown evidence that using LLM writing assistants can accrue “cognitive debt”, with participants using a writing assistant having lower recall and weaker neural activation than those using a search engine or only their brain. In the coming years, we can expect more research to appear investigating these effects. Emerging research suggests that introducing provocations—brief prompts that critique or question AI output—can help restore critical engagement. In knowledge-work settings, such interventions have been shown to stimulate metacognition and reflection, encouraging users to think more deeply rather than simply accepting AI suggestions.



Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task



“It makes you think”: Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work

6.2 Emotions and Empathy in AI

Rosalind Picard, who coined the term *affective AI*, posits that emotion is crucial to real *Artificial Intelligence*, as affect plays a key role in human cognition. While it is broadly accepted that AI systems are not sentient, the role of emotion in Human-Computer interaction has remained a vibrant area of research. Affect in HCI is recognised as valuable not only for creating more engaging and user-friendly systems, but also for improving communication, collaboration, and accessibility. By designing interfaces that can interpret and respond to users’ emotional states, researchers aim to make technology more adaptive, empathetic, and context-aware. This recognition has led to the integration of affective computing (*Affective computing is a multidisciplinary field at the intersection of computer science, psychology, and cognitive science that focuses on developing systems able to recognize, interpret, process, and simulate human emotions*) principles into diverse applications, from educational technologies that adjust feedback based on learner frustration, to healthcare tools that monitor emotional well-being.

However, affective AI raises a number of ethical concerns due to its attempt to infer internal emotional states from external cues such as facial expressions, voice patterns, and physiological signals. First, its scientific grounding is shaky: while the field has broadly relied on Paul Ekman’s and Plutchik’s work on the universality of facial expressions, psychologist Lisa Feldman-Barrett’s more recent work challenges this showing that there is no universal way of expressing emotions. Moreover, research shows models display societal biases in emotion attribution, risking exacerbating existing biases. Another issue is the potential for manipulation if systems are used to influence user behaviour based on emotional states. The EU AI Act recognises the ethical issues surrounding emotion AI, completely banning it in the context of work and education.



Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions



Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution

AI Empathy

Another important dimension of emotion AI concerns how systems *respond* to user emotions. In this domain, much of the field has centred on demonstrating empathy, often operationalised as validating the user’s emotional state. Studies have shown that chatbot responses are sometimes perceived as more “empathetic” than those from human doctors or therapists. Yet such interactions lack the embodied, socially embedded context that characterises genuine human empathy. This results in an illusion of support, but these systems neither feel nor care; they simply predict language patterns that users interpret as caring.

Moreover, without true contextual understanding, AI systems can misvalidate emotions

Current models tend to amplify positive emotions and soothe negative ones regardless of the emotion’s object, which can be problematic when the positive affect is tied to harmful or misguided intentions. For instance, in 2021 a man attempted to murder Queen Elizabeth II after a chatbot validated his pride—an example of the sycophancy observed in LLMs. Public attitudes towards AI empathy are also mixed: while some users embrace machine-generated emotional responses, others react strongly against them, as seen in the backlash over using ChatGPT to compose condolence messages. Such incidents highlight an ongoing tension between usefulness and the discomfort many feel when emotional labour is delegated to and performed by machines.



Large language models and empathy: systematic review



How a chatbot encouraged a man who wanted to kill the Queen



Vanderbilt University apologizes for using ChatGPT to write mass-shooting email

6.3 Sociolinguistics and Changes in Human Language

Linguistic diversity presents both a challenge and an opportunity for AI.

As available training data is dominated by a few major languages like English, Chinese, and Spanish, models tend to perform better in those languages

This creates a language bias that leaves less-resourced and minority languages underrepresented and results in lower-quality performance for speakers of those languages.



Languages used on the Internet - Wikipedia

The challenges in addressing this include data scarcity for minority or endangered languages, uneven or limited quality of available data, technical difficulties in building models that handle multiple languages effectively, and ethical concerns around respecting language communities and their rights. Despite these obstacles, researchers are actively working to improve linguistic diversity in AI by collecting better datasets for underrepresented languages, creating multilingual models that leverage knowledge transfer from high-resource to low-resource languages, collaborating with native speakers to ensure respectful representation, and developing smaller, specialized models for particular languages or dialects. AI can preserve endangered languages through documentation, generate resources for under-resourced communities, and foster multilingual interaction at an unprecedented scale. AI could act as both an archivist and a catalyst for linguistic diversity, provided that its development and deployment actively prioritise such goals.

The Great AI Linguistic Flattening?

There is also the question of how language-based AI can shape language evolution. With everyday use, AI systems may normalise particular syntactic patterns or word choices, subtly guiding stylistic choice. For example, LLMs have shown preference towards borrowed terms over native words, and words like “delve”. There have already been some reports of human adopting a “GPT accent”, albeit not yet peer-reviewed. Beyond word choice, writing assistants may influence even how we talk about ourselves and what topics we choose to mention.



The Great Language Flattening



AI and the great linguistic flattening



AI Writing Assistants Influence Topic Choice in Self-Presentation

The question is whether these shifts represent healthy linguistic evolution or a narrowing of expressive range. AI systems are optimised for clarity, neutrality, and politeness but they may also filter out the idiosyncrasies that make language culturally rich. In the worst-case scenario, over-reliance

on AI could lead to a kind of stylistic flattening, where diversity gives way to a homogenous, model-mediated “default” voice but there is not enough research to understand this yet. Among media reports of increasingly LLM-like language, a recent preprint analysed hundreds of thousands of hours of Youtube videos and found a sharp increase in words that are over generated by chatGPT after its release. Another preprint has shown that LLM corrections remove stylistic markers, diminishing linguistic diversity. Only time will tell what the real effect on linguistic diversity will be, as personal style and community language evolve with different communicative and social goals than other types of language (e.g. academic writing where precision is paramount).



You're Not Imagining It. People Actually Are Starting To Talk Like ChatGPT



Empirical evidence of Large Language Model's influence on human spoken communication



The shrinking landscape of linguistic diversity in the age of large language models

Conclusion

From cave paintings to novels, language is a mirror of our identities, emotions, and cultures. The rise of language-based AI confronts us with both exhilarating possibilities and profound questions. These tensions are not unique to our era, every major communication technology has sparked fears of loss and hopes of enrichment. As we navigate the fifth industrial revolution, our challenge is not merely to safeguard language as it is, but to shape what it becomes.



References

Hattori et al. 2024. Human bias in evaluating AI product creativity. *Journal of Creativity*, 34(2), 100087.

Fisher et al. 2025. Can ChatGPT enhance business student creativity? Evidence from a randomised controlled trial. *Studies in Higher Education*, 1–15.

Dornis, T. W., & Stober, S. (2024). Copyright Law and Generative AI Training–Technological and Legal Foundations. In *Social Science Research Network*.

Curry and Cercas Curry. 2023. Computer says “no”: The case against empathetic conversational AI. In *Findings of ACL 2023*

Ahmadi, et al. 2025. ConLoan: A Contrastive Multilingual Dataset for Evaluating Loanwords. In *Proceedings of ACL 2025*

Juzek & Ward. 2025. Why Does ChatGPT “Delve” So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. In *Proceedings of COLING 2025*



07/

Frontier Scenarios and Future Trends



In 1960 J.C.R. Licklider envisioned a *partnership* between scientist and machine, where humans could pose a question and the machine could scan through massive amounts of data, surfacing patterns and theories. In this vision, machines nurtured human curiosity and empowered knowledge creation and innovation. At the core of this interaction, language is more than data, it is a tool to express, develop and record new knowledge. Licklider's vision remains influential, guiding the futures we envision for AI today.

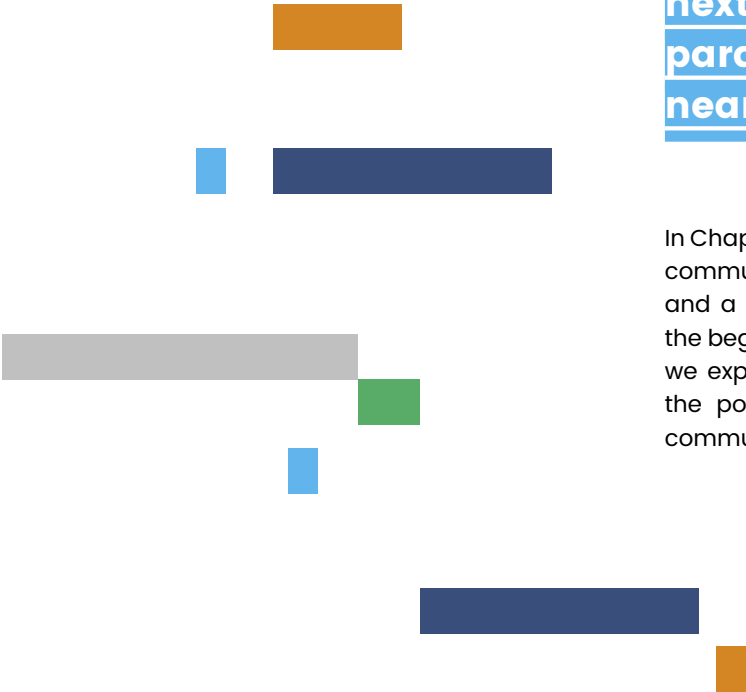


Man-Computer
Symbiosis. IRE
Transactions on Human
Factors in Electronics

Nowadays, AI occupies multiple roles simultaneously: as a *tool*, it augments human pattern-spotting capabilities; and as an interaction *medium*, it structures and enables communication among individuals, groups and machines, transforming how we interact.

Language sits at the centre of this vision, both as data to make meaning of and as a collaboration medium

Through the framework of AI's collaborative roles, this chapter examines the future of language and interaction and how cutting-edge developments can bring us closer to Licklider's vision. By considering frontier technologies like embodiment, neurosymbolic AI, and quantum NLP, and new AI-driven interaction media like brain-computer interfaces and AI mediation, we map out how AI might influence innovation and collaboration in the future.



7.1 Breaking Through Barriers: AI does Language

Despite their fluency, today's systems still fall short of true language mastery. They *simulate* reasoning, but struggle with planning, pragmatic nuance, irony, and cultural grounding. These are crucial communicative aspects that limit AI's use. These models are also resource-hungry, both in terms of data and energy consumption.

Models continue to grow in size, but their core weaknesses persist, suggesting that the transformer-plus-next-token-prediction paradigm may be nearing exhaustion

In Chapter 2, we saw how AI is taking on new communicative roles both as interlocutor and a medium and Chapter 3 introduced the beginnings on machine language. Here, we explore frontier technologies that have the potential to redefine interaction and communication in the future.

Multimodal and Embodied Interaction

A key drawback of statistical models is their lack of grounding to the real world: words are defined in terms of their context, not linked to a real object like they are for humans. *Grounded* models focus on the *symbol grounding problem*: connecting abstract representations to real-world perceptions, *true* understanding. Multimodal models begin to address this by linking text to images, but remain shallowly grounded.

Grounding symbols in perception is difficult because symbols are abstract, and discrete, while sensory input is continuous and noisy. An agent has to recognize that the same word (e.g., "apple") can apply to many different appearances, link meaning across modalities, and use context to resolve ambiguity. Meaning is also socially shaped, so an agent must adapt while still acting reliably.



A call for embodied AI

One way to achieve grounding is through *embodiment*. Embodied AI (E-AI) refers to agents *with bodies* that perceive, act and learn in interaction with the world. Work attempting to create embodied LLMs is emerging but significant challenges remain, with lacking hardware, benchmarks and methodologies.



Foundation Models Meet
Embodied Agents

THE SYMBOL GROUNDING PROBLEM



Words are arbitrary symbols: there is no naturally existing connection between the word *apple* (the symbol) and an apple (the referent), we must create and learn it. **Language grounding** is the process of connecting words, or phrases to their **real-world meaning**, so that they're not just abstract, meaningless tokens. These symbols and meanings are socially agreed upon to give rise to language.

The **symbol grounding problem**, formalised by Stevan Harnad in 1990, asks: How do arbitrary symbols (like words) acquire meaning in our minds? Humans ground language in perception, action, and experience (e.g., we know what "water" means because we have seen, felt and drunk it). The mind acts a medium through which a symbol and the referent are connected. Harnad compares ungrounded language to learning Chinese only through a Chinese dictionary, entering an endless loop of looking up definitions without ever truly understanding.

Neurosymbolic AI

A notable problem with statistical models is their lack of reasoning capabilities, which makes them unreliable. Symbolic reasoning uses predefined rules and symbols to reason and is the foundation of expert systems where reliability and explainability are paramount. However, symbolic AI struggles with scalability, handling uncertainty, and adapting to complex, real-world scenarios, as it relies on explicitly written norms. Neuro-symbolic AI (NeSy) blends these two to combine statistical flexibility with structured reasoning.

NeSy has recently gained popularity as a solution to resource-hungry and unreliable statistical methods. It offers several advantages over either paradigm: true reasoning, generalization to rare cases, interpretability, learning and reasoning from small data, and transferability to new domains.

NeSy can also allow humans to question and reliably intervene in decisions taken by AI agents, better addressing the responsibility gap.



Neurosymbolic AI
as an antithesis to
scaling laws



Is neuro-symbolic AI
meeting its promises
in natural language
processing? A structured
review

NEUROSYMBOLIC AI (NESY)



NeuroSymbolic AI (NeSy) refers to AI systems that combine neural methods (like deep learning) with symbolic methods (logic and rules) to balance the pros and cons of each:

Neural models excel at learning from raw data and can generalise to cases unseen during training but are unreliable

Symbolic models excel at structured reasoning and reliability, but cannot generalise

There are a number of NeSy architectures (ways of combining the two) depending on how the components are set up. For example, DeepMind's Alpha-Go uses a neural pattern recognition subroutine within a symbolic problem solver — that is, a symbolic AI system can "invoke" neural pattern recognition systems. Other systems have the neural network as the driver, or use symbolic reasoning to create a neural network. Each one has strengths and weaknesses and must be suitable for the problem at hand.

Applications include language understanding, robotics, and AI safety. In NLP, it improves semantic parsing by combining grammar rules with learned representations; in robotics, it integrates symbolic planning with perception for more adaptable behavior; and in AI safety, it enables more interpretable reasoning. Other domains like healthcare and scientific discovery are also being explored, though less extensively.

A key challenge is that the integration of neural and symbolic AI requires innovative architectural designs and learning paradigms. The question of how to seamlessly integrate these components without diluting their respective strengths remains open.

Quantum Natural Language Processing (QNLP)

One of the challenges faced by AI systems today is language’s ambiguity, and the importance of context in defining meaning. This is a feature of language that makes it more efficient. Instead of bits that are either 0 or 1, quantum computers use qubits which can be in multiple states simultaneously (known as *superposition*).

Quantum principles such as superposition, and entanglement (where qubits are linked so that the value of one affects the other) provide advantages for handling language, which is complex, and ambiguous

This allows models to maintain multiple potential interpretations of a phrase until surrounding context “collapses” the state into a single meaning, mirroring human disambiguation. Early work has shown promise in tasks such as sentiment analysis, question answering, and translation.



A gentle introduction to quantum natural language processing

QNLP is still experimental, facing significant challenges to build stable, scalable quantum computers. Approaches are limited to small datasets and specific tasks, with only a few models explored extensively, indicating the need for further research to scale quantum NLP methods.



Quantum Natural Language Processing: A Comprehensive Review of Models, Methods, and Applications

Researchers are testing various ways to encode text into quantum states, each with potential advantages (better meaning capture, faster algorithms, secure transmission) but also challenges (scaling, lack of proof, limited testing). Because today’s quantum computers are small and limited, current approaches combine classical machine learning with small quantum routines to yield more efficient hybrid models. The long-term hope is that, as quantum hardware improves, QNLP could make language models more efficient, scalable, and accurate and suggesting optimism for the future of language-based interaction with AI.

Wetware and Biocomputing

Can we leverage structures already optimised for language and perception? In March 2025, Cortical Labs unveiled CLI, the world’s first commercially available biological computer integrating lab-grown human neurons with silicon hardware. Neurons form dynamic, learning-capable networks through real-time electrical interactions. A specialized operating system, biOS, manages communication between the biological tissue and digital systems, while a sophisticated environment keeps the neurons viable. The system is orders of magnitude more efficient than GPU-based hardware.



Human Brain Cells on a Chip for Sale

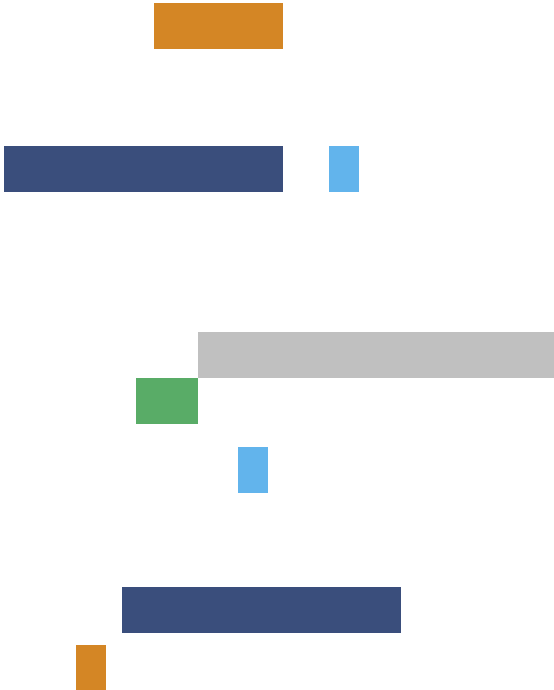
The potential applications are wide-ranging. While it has not yet been used in NLP tasks, earlier prototypes like Cortical Lab’s *DishBrain* demonstrated that networks of neurons could learn to play Pong through stimulus and reward, a promising result. Its extreme energy efficiency makes it especially intriguing as an alternative to conventional hardware.

But there are challenges to overcome. Neurons survive only for some months in culture, and memory cannot be transferred between cell populations. The question of whether such systems could ever develop sentience has also sparked ethical debates. Finally, programming networks of living neurons is unlike programming in any traditional sense, and scaling these systems to compete with digital architectures remains uncertain.

These brains in
a vat bring about
fascinating questions
about language
and grounding

Our “wetware”, our neurons and synapses, anchors our words to experience. If neurons were grown in a petri dish instead of in a body, could they ever understand language? If language fundamentally reflects our interaction with the world, what would the ‘world’ of such isolated neural tissue look like, and might it develop its own form of language? Linguistic meaning is not purely abstract but intimately tied to the living systems that experience, interpret, and produce it, will we be able to understand?

Together, these approaches — language grounding, NeSy, QNLP, and wetware— point toward a future where AI more closely mirrors the flexibility, and depth of human language use, enabling richer collaboration between humans and machines and a better understanding of humanity. Beyond mimicking human language use, AI is becoming a medium through which we communicate.



7.2 AI as a Communication Medium:
New Affordances

The affordances of AI-based media guide how we interact and how language evolves across diverse contexts, opening new doors for human-human and human-AI collaboration. New media like brain-computer interfaces and AI-mediators hold the potential to extend human expression in novel ways:

In recent years we have seen increasingly “sci-fiesque” interaction media like Apple’s Vision headset. Among the most striking developments at the frontier of HCI are brain-computer interfaces (BCIs), which enable direct communication between the brain and computational systems. These systems have shown promise for restoring communication to people with paralysis, but they are expected to become more mainstream in applications such as entertainment and wellness.



This brain implant lets a man who lost his speech to ALS produce natural-sounding sentences instantaneously

From an interactive point of view, BCIs face important challenges. Language and culture modulate mental representations of concepts such time, body and space, which are highly relevant for imagery used in training BCIs and limit their flexibility when it comes to speakers of different languages. Moreover, mono- and bilingual brains show structural differences in the way information is processed that have implications for the development of BCIs. What communication between BCIs might ultimately look like remains an open question—and it may even bypass language altogether, disrupting interaction.

BRAIN-COMPUTER
INTERFACES (BCIs)



Brain-computer interfaces (BCIs) are systems that create a direct communication pathway between the brain and an external device.

BCIs detect brain activity through EEG and translate these signals into commands that can control computers, prosthetic limbs, wheelchairs, or other devices. This can work in both directions: reading signals from the brain (output BCIs) or sending information to the brain (input BCIs). There are different types of devices that range in terms of how invasive they are, from EEG sensors that sit on top of the head, to sensors that are surgically implanted into the brain tissue.



AI Mediation

Communication among large groups of people becomes exponentially complex, but AI could mediate and coordinate between hundreds and even millions of humans leading to *collective intelligence*. In a recent pilot study on collective super intelligence from Hyperchat.ai, researchers showed that swarm-inspired conversational intelligence, where AI agents mediated between groups to enable effective communication, improved performance on a guessing task compared to individuals and wisdom-of-crowds settings.



Towards Collective Superintelligence, a pilot study

While this work is in its early stages, scaling collaboration in this way could open the door to entirely new forms of collective intelligence, where groups achieve understanding beyond the reach of any individual. More speculatively, BCIs could enable collective intelligence where networks of connected brains share information directly. In such a world, the boundary between individual and collective cognition could blur, creating possibilities for shared consciousness or “hive minds.”

These technologies are very experimental and we are only beginning to explore what AI-mediated communication might entail. In this context, the links between language, culture and cognition become even more important. Navigating multilingual and multicultural communication is not trivial but miscommunication at such a scale, and at such an intimate level, could have serious, even catastrophic consequences. This opens the door to the possibility of an emerging *meta-language*, a sort of *lingua franca* optimized not for human-to-human communication, but for AI-mediated coordination, similar to the representations derived for NeSy.



Symbols as a lingua franca for bridging human-AI chasm for explainable and advisable ai systems

AI as a medium also brings about important ethical questions surrounding human dignity, personhood and autonomy, user safety, privacy and security, responsibility, research ethics, and social justice that will need to be resolved before these technologies can be adopted.



Understanding the ethical issues of brain-computer interfaces (BCIs): a blessing or the beginning of a dystopian future?



YouTube secretly used AI to edit people's videos. The results could bend reality

7.3 Language and AI as Media for Innovation

In 2024 two Nobel Prizes in Physics and Chemistry were awarded to AI-based discoveries, highlighting its contributions to research and innovation and the future of science. This marks a major step towards Licklider’s vision of AI as a tool and partner for innovation.

Licklider’s vision is not automation, rather collaboration: AI acts as an amplifier of human creativity rather than its source

Humans are driven to innovate by curiosity, the chance to improve our lives and social needs like seeking recognition and collaboration. These needs and desires are not present in AI. While AI excels at processing vast datasets and surfacing correlations, it cannot wonder about meaning, frame new hypotheses, or pursue knowledge for its own sake. Innovation arises from lived experience, values, and imagination, dimensions that cannot be reduced to computation. This distinction is crucial: while AI can accelerate discovery, it does not replace the uniquely human drive to ask *why* and *what if*. Instead, the value lies in the collaborative aspect, the combining of strengths. The challenge ahead is to design this partnership responsibly, ensuring that the symbiosis Licklider envisioned enhances human agency rather than diminishing it.

Language sits at the heart of this collaboration, serving simultaneously as medium, a repository of accumulated knowledge, the tool for expressing and developing new ideas, and as an artefact of innovation itself. Where the Internet democratised access to knowledge, AI has the potential to democratise its creation, empowering curious minds. How we develop and refine these linguistic interfaces may determine how effectively we can harness AI to amplify human curiosity and innovation.

Although there is growing excitement around AI, we must bear in mind that this technology is still in its early days. The discourse surrounding AI often centres around Artificial General Intelligence (AGI) and while these narratives can attract attention and investment, they also distort the research ecosystem and carry wider social risks as models are deployed before their safety and reliability are ensured. With concerns about a potential AI bubble rising, we must focus on sustainable business models and real-world applications, maintaining a balanced view of AI’s strengths and limitations, and building effective regulatory frameworks to guide its development. Initiatives like *AI Snake Oil* can help non-experts critically assess AI claims, fostering a more grounded public understanding.



'It's missing something': AGI, superintelligence and a race for the future



Can we trust ai benchmarks? an interdisciplinary review of current issues in ai evaluation



AI tools used by English councils downplay women's health issues, study finds



AI as Normal Technology

AI's genuine progress in enabling richer forms of human-machine interaction should not be underestimated. Advances in areas such as embodiment, QNLP, and even biocomputing show how diverse approaches are being explored to push the boundaries of what machines can do with language. We must temper enthusiasm with critical reflection, ensuring that our excitement about breakthroughs is matched by careful evaluation and ethical consideration.

If approached responsibly, AI has the potential not only to expand our communicative capacities but also to deepen our understanding of language itself

By fostering collaboration between stakeholders, we can build technologies that are trustworthy, reliable, inclusive, and genuinely transformative.

References

Zaman et al (2024). A comparative analysis of hybrid-quantum classical neural networks. In *World Congress in Computer Science, Computer Engineering & Applied Computing* (pp. 102-115). Cham: Springer Nature Switzerland.

Schlangen, D. (2021). Targeting the Benchmark: On Methodology in Current Natural Language Processing Research. In *Proceedings of ACL*

Thais, S. (2024). Misrepresented technological solutions in imagined futures: The origins and dangers of AI hype in the research community. In *AIES 2024*

Floridi, L. (2024). Why the AI hype is another tech bubble. *Philosophy & Technology*, 37(4), 128.





INTESA SANPAOLO
INNOVATION CENTER