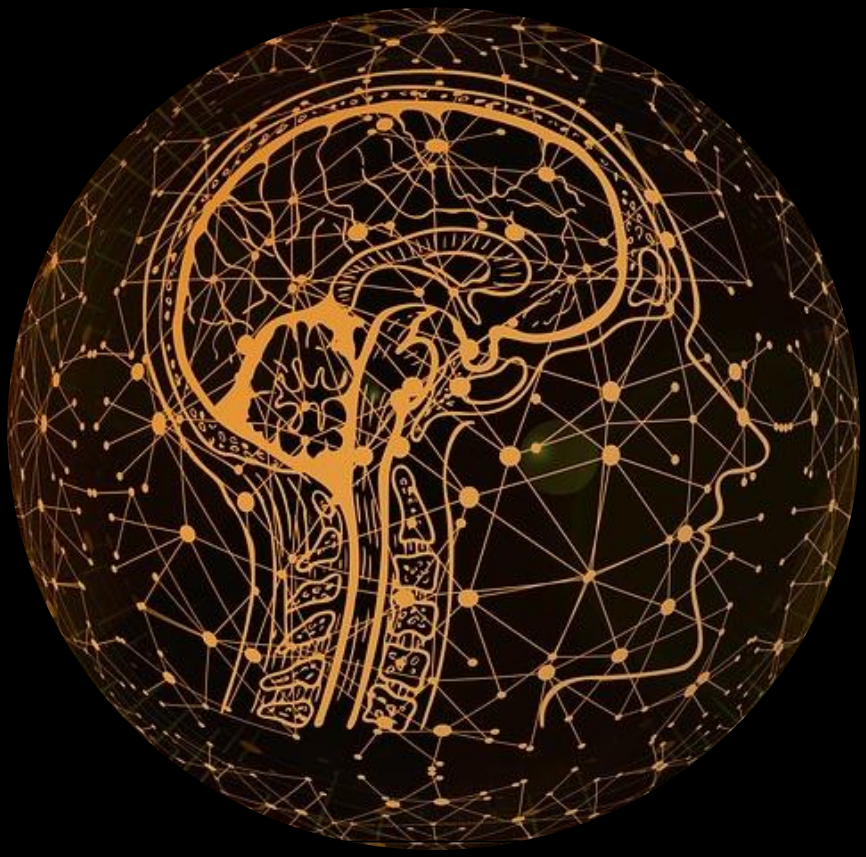
nhnai

New Humanism in the time of Neuroscience and Artificial Intelligence



NHNAI white paper: Societal exploration of humanism at the service of AI ethics and governance

First version - January 2026

UCLy LYON CATHOLIC UNIVERSITY
UNITÉ DE RECHERCHE confluence SCIENCES ET HUMANITÉS EA1598
IFCU International Federation of Catholic Universities

New Humanism in the time of Neuroscience and Artificial Intelligence

NHNAI white paper: Societal exploration of humanism at the service of AI ethics and governance

This document constitutes a first version elaborated by NHNAI coordination team thanks to the direct the support and inputs of NHNAI partners. It is now published and opened for comments and reactions. Feedback will lead to the development of a second updated version.

Preamble

Humanity is confronted with major challenges including climate change, inequalities, geostrategic tensions, weakening of democratic political organizations (notably with acute questions about the equilibrium between public and private powers as well as about collective intelligence and the threat of post-truth). Moreover, these multiple challenges all occur at the same time and with very rapid and brutal dynamics. As illustrated with the well-known analogy with the pharmakon, AI has considerable potential either to support mitigating these challenges or to amplify them.

The current era of AI—driven by advances like deep learning and Large Language Models (LLMs)—is marked by frenetic development and intense global competition. This rush is fueled by the perception of AI as a primary economic driver, leading to a "tyranny of tardiness" where attempts at regulation are protested as economic suicide. This hectic, market-driven environment, coupled with growing power asymmetries favoring tech giants who often adhere to techno-solutionist ideologies, makes ethical AI development extremely challenging.

The NHNAI project builds upon the core idea that, to move beyond this uncritical embrace of technology and put AI genuinely at the **service of humanity**, collective and intense effort of ethical capacity-building must be conducted to empower the individual and collective discernment and contribute to a strong horizontal and bottom-up support to the **democratic governance of AI**. Merely pointing out power imbalances isn't enough. They must be confronted. But we believe that adequately orientating the development and use of AI is the responsibility of all concerned persons. In fact, power asymmetries are not the sole obstacle on the road of robust democratic governance of AI. Another challenge is to be able to set goals and purposes to AI. This is a decisive collective responsibility. What are the society project AI should serve? What are the visions of human nature, development and flourishing that will operate in the background?

The present white-paper results from NHNAI network's effort to contribute to this necessary collective endeavor of discernment. It proposes several recommendations emitted by the academic experts of the network (notably based on the various discussions in the nine countries that participated in the first 2022-2025 phase). Recommendations are divided into three main components: 1) recommendations for the organization of collective reflection to build strong support to democratic governance of AI, 2) recommendations on basic elements of understanding of AI as well as of what it means to be human (elements without which collective discussions could be impaired), 3) recommendations on important topics that should be explored in collective reflection.

Before proposing a detailed analysis that deploys the full content of NHNAI recommendations, the document starts with an executive summary with the recommendations presented in a condensed form. The reader can either consult this summary and refer occasionally to the detailed analysis for more details, or directly begin with the detailed analysis.

Table of content

Executive Summary

Detailed analysis

Executive summary

Key inspirations

The collective discernment required for ethical AI governance must fundamentally shift its spirit and focus, moving away from putting humans and AI systems in direct competition. We must recognize that replacing a person with AI is **never "changing nothing."** This decision always substitutes the richness of a living person's lived experience with pure mechanism and automatism. Therefore, the best way forward is not to focus solely on maximizing efficiency with AI in roles where humans might seem faulty or less performant.

This calls for a necessary collective effort of discernment, demanding considerable political commitment from human communities. We must change the central question guiding AI development: instead of focusing only on "*What can AI do better than humans?*" we should primarily wonder: **"How can AI technology support us in becoming better humans?"** The question should not be "*What is our place as humans in the new world of AI?*"—as if we were merely adapting to a force beyond our control—but rather: **"What is the place of AI technology in our human world, and what contribution can it bring to the development of more humane societies?"** Framed this way, it becomes clearly visible that AI development is fundamentally a political and ethical issue, not just a technical one.

Crucially, we face a **circular problem** where developed AI systems and uncritically adopted uses can undermine the very capabilities needed for proper discernment: critical thinking, free attention time, strong decision-making, and sensitivity to life.

This makes collective commitment to discernment efforts even more crucial. It is legitimate to be enthusiastic and optimistic with AI. We must be so. AI comes with tremendous potential to support us in our development and flourishing. But this potential cannot be fully realized without the involvement of societal communities, engaged in profound efforts to nurture discernment capabilities, notably by fostering better understanding of AI technology themselves (AI literacy) as well as of human nature and condition. No doubt AI can make wonderful contributions to human flourishing, but only if we always foster and keep clear awareness of the price we are called to pay in terms of commitment to challenging efforts of ethical and political discernment.

A collective societal reflection to build strong support for democratic governance of AI

Recommendation 1: An inclusive and horizontal reflection

To ensure effective and robust **AI regulation**, society must establish an **inclusive and horizontal reflection** process. While expert groups and political representatives are necessary, delegating the entire regulatory effort to them risks a technocratic or elitist setting, which may not yield the best outcomes.

Genuine governance requires mobilizing **all concerned stakeholders**. Broad participation is vital not only for the **quality and content** of the regulations but also for ensuring **high social awareness** of these rules. Without widespread understanding and involvement, the **enforcement** and long-term success of AI regulatory devices will be severely undermined.

Recommendation 2: A reflection not only on uses of AI systems but also on their design

The governance of Artificial Intelligence must fundamentally move beyond regulating mere **uses** to encompass the **design and development** phases of AI systems. The traditional cultural view, inherited from modernity, often separates **neutral facts** (technology/science) from **values** (societal/democratic will), leading to the simplistic belief that ethics only applies when choosing how to use a ready-made technology.

This view is profoundly flawed, as a long-standing techno-critic tradition shows that artifacts inherently embed **political and ethical implications**. Especially with AI, ethical considerations begin well **upstream** of usage. For instance, concerns over biased classification tools, algorithmic decision systems, or recommendation engines driven by attention capture demonstrate that the ethical "quality" of outcomes is largely **built into the system's design**, not just determined by user choice.

Therefore, the background reflection society must conduct is one that recognizes the need for ethical deliberation to happen **upstream**—at the **design and development stages** of AI. This requires participatory input from citizens and stakeholders, not just to inform external legislation, but to embed values directly into the creation process through methodologies like **Value Sensitive Design (VSD)**. Only by understanding that ethics and governance must permeate the entire AI lifecycle can society provide the necessary foundation for robust and meaningful regulation.

Recommendation 3: A transformative effort of genuine political reflection, with the support of experts

Effective AI ethics and governance must embrace a **transformative political reflection** supported by experts, moving far beyond simply collecting public opinions. The design of AI systems possesses an **ineliminable political dimension** that cannot reduce to the mere juxtaposition of already existing opinions that mere surveys could capture or to the opinion of the majority in case there are some conflicting trends. This would be assuming either that valid answers and solutions are available, ready-made to be operationalized, or (in a more pessimistic way) that nothing more can be done than such a collection. This is particularly problematic since current societal preferences often contribute to the very problems AI raises (e.g., biased algorithms reflecting societal biases, or uncritical adoption of disruptive economic models).

Addressing the complex issues AI ethics and governance raise, which are often **'wicked problems'** blending technical, ethical, and political questions, necessitates **social**

transformations. The approach requires **transdisciplinary co-production of solutions** where diverse experts (philosophers, computer scientists, economists, etc.) support the collective societal reflection. Crucially, this support cannot be a technocratic exercise. Experts are not expected to bring answers to issues at hand. Instead, their expertise must facilitate processes of **self-criticism, mutual enrichment, and self-transformation** within social groups, enabling them to build robust societal inputs for the orientation of AI development and uses.

This collective reflection effort must challenge the misleading **fact-value dichotomy**, which wrongly treats ethical or political questions as purely subjective matters of free will. By engaging in **communities of rational discussion and deliberation**, participants in the collective effort of reflection fulfill the duty to sincerely seek validity and truth, moving beyond a simple patchwork of conflicting opinions. This rational, collective reflection is essential to establish shared understandings of issues at hand, making room for legitimate tensions (like privacy vs. security), building a strong ground for effectively guiding AI development and uses toward agreed-upon ethical goals.

Recommendation 4: Ethical capacity-building for bottom-up and horizontal guiding forces

Robust AI ethics and governance necessitate **building strong ethical capabilities** within societal communities. This collective, reflexive, and transformative work (recommendation 3)—mapping issues and seeking answers—empowers citizens and stakeholders to contribute to AI guidance.

This **empowerment** is crucial for several reasons:

- It permits a bottom-up contribution to the design of **top-down** ethical principles and legal regulations.
- It facilitates the **contextualization and enforcement** of rules locally, creating **bottom-up and horizontal forces** of orientation.
- It fosters enlightened daily and consumer choices, contributing to creating **viable economic space for ethical entrepreneurship**.
- It enables communities to define **high-added value use cases**, ensuring AI technologies genuinely contribute to human societies.

Widespread ethical capacity-building is essential to guide AI development effectively.

Recommendation 5: A dedicated exploration on what being human means

Effective **AI ethics and governance** critically depend on a foundational reflection on **humanism**—specifically, what it means to be human, and what we *want* to be as humans. In fact, many major ethical principles for AI—such as keeping the human **"in the loop"** or aiming for **"human flourishing"**—directly appeal to the idea of the human. More fundamentally, the very nature of ethics, defined by figures like Ricoeur as the aim of the **"good life" with and for others in just institutions,** requires reflecting on what it means to be human to forge the meaning of those terms.

The rise of technologies like AI, which provides humans with the power of **in-depth self-modification** (anthropotechnè), makes a reliable compass about our human nature more necessary than ever.

However, the notion of humanism is neither clear nor consensual, facing criticism from anti-, post-, and trans-humanist currents that notably question its problematic emphasis on autonomy or its role in the "master and possessor of nature" myth. Instead of outright rejection, the recommendation is to engage in a **renewed exploration** of humanism. By critically re-engaging with fruitful ideas—like Kant's central focus on **freedom, responsibility, and the faculty of judgment**—and coupling them with insights from AI and cognitive science, we can foster a **collective exploration** to outline the contours of a **new humanism**. This deepened, shared understanding of our own human nature (of who we are as well as of who we should be) is indispensable for developing reliable ethical guidance for AI.

Fundamental milestones on AI and humans to support the societal reflection

Recommendation 6: Ensuring a robust and empowering understanding of AI technologies

Currently, AI is often misrepresented as a limitless, inexorable wave leading toward AGI or SAI, fueling narratives of human obsolescence or lost control. These misleading representations ignore AI's crucial dimension as a **human artifact**, screening off its political and ethical reality. Therefore, it is essential to secure a not-too-abstract understanding of AI technologies within participatory communities. This literacy is the necessary precondition for citizens and stakeholders to effectively build, deliberate upon, and guide the positive societal projects AI should serve.

Demystifying machine learning

Machine Learning (ML) is a sophisticated, technical search for the optimal **parametrization of a computational architecture** to execute tasks that resist conventional, step-by-step programming (like complex image classification). Engineers design a specific architecture with many different types of operations and arrangements of them that are specified by free parameters (e.g., coefficients). They then write a trial-and-error (or similar) program to adjust these parameters, guided by explicitly defined **feedback**.

This minimal level of explanation is **crucial for demystifying AI and managing expectations**, as it reveals there is **not a unique, big, all-powerful AI**. Instead, there are **various techniques** highly dependent on the architecture (like convolutional nets or transformers) and, most importantly, the nature of the feedback provided:

1. **Novelty and New Ways of Solving:** In narrow, specialized configurations, like board games, video games or other simulated environment, the feedback can be defined mathematically. This allows the system to possibly find genuinely **novel solutions** and

better ways of achieving the goal represented by the feedback. Here, systems *can* produce results humans were unaware of, but these are niche contexts.

2. **Reproduction and Limited Generalization:** In the most widespread applications (like LLMs and image classification), the feedback is defined by comparison with a large dataset of **examples**. In these cases, the system's primary goal is to imitate and correlate, not innovate. It is **misleading to expect radically new results**; the system is rewarded for reproducing the past, not inventing the future. Therefore, the power of such systems to generalize over new problems is very limited.

Furthermore, this a bit de-abstracted view permits better anticipating the reliability of the different systems. It allows mitigating the widespread tendency to describe LLMs and other (deep) machine learning programs as inscrutable black boxes. In fact, operations such programs do are known. It is in principle possible to look at the values of the various free parameters and the calculations they lead to. The problem is that this is poorly informative on why a given program works fine or not. We often lack a reliable **theory of error**, resulting in the possibility of unpredictable "hallucinations." But, in these matters of reliability again various types of systems can be distinguished. For instance, reliability may be reasonably evaluated and thus expected in specialized, empirically testable tools (for instance a program specialized in medical image classification), but our expectations should diminish drastically in general-purpose generative AI, the usages of which should be adapted in consequence.

Recalling the materiality of AI

AI and digital technology are often misleadingly pictured as immaterial. However, all programs run on physical **hardware** designed for **automatic, mechanical transformation** of material configurations (e.g., magnetic orientations, electronic states) to which humans have assigned meaning (like 0s and 1s, words, or numbers). In this regard, the computer can be seen as the culmination of a long history of information technologies, dating back to the very invention of **writing** (which precisely consists in giving meanings to particular material shapes and is the very first step to afterward build automata that will act upon these shapes).

A fundamental property of the computer is its **intended, undeviating inertia**. It processes information by efficiently and precisely manipulating these material configurations according to a program. Computers do not inherently contain meaning, emotions, or consciousness; they are simply **fantastic machines** that act upon configurations that mean something for us to mechanically create new configurations, which *we* then interpret as text and images (possibly expressing feelings).

Highlighting the direct dependence on human intelligence

It is essential to **marvel at the successes of AI for the right reasons**: they demonstrate humanity's ability to build inert, complex mechanisms that *simulate* intelligent behavior. AI is fundamentally a product of, and **irreducibly dependent** upon, human intelligence.

The notion of a magical Super Artificial Intelligence (SAI) that produces true outcomes we cannot verify is illusory. Humans remain completely in charge of building the systems and assessing their results. High levels of human intelligence are required across the entire pipeline:

- **Design:** It asks for smart programmers and engineers to create computational architectures and learning procedures.
- **Guidance:** It requires domain experts to formally frame the feedback mechanisms, and-or many intelligent humans to provide good examples for the training datasets (e.g., journalists, researchers, any person sharing some content on information systems).
- **Assessment:** Human effort and intelligence are needed to build trust, assess quality, and define the systems' reliability limits. One must acknowledge the existence of a genuine 'control' problem when reliability of systems is not warranted enough (especially with powerful generative AI systems whose outputs might become very difficult to anticipate and secure).

The **highest level of collective human intelligence** is required for the crucial task of defining adequate **orientations and goals** for AI. It will also demand a lot of human strength and intelligence to refrain from using, in not secured enough settings, systems that would not present sufficient warranties of reliability.

Recommendation 7: Securing some basic intuitions on the specificities of humans by comparison with machines

Effective AI ethics and governance require a reflection on human specificities compared to machines. Disruptive or prophetic claims suggesting machines could (soon) possess core human traits like consciousness or free will should be taken with extreme caution. While blind dogmatism should be resisted, it is far from obvious that there currently are good reasons to substantially revise our basic intuitions upon the specificities of humans by comparison with machines.

Resisting the injunction to "outcomism"

Communities engaged in the reflection upon AI ethics and governance must actively **resist "outcomism,"** the prescription to restrict the discussion of AI's human-like traits (consciousness, intentionality, free will) solely to comparing observable **outcomes** between AI systems and humans.

This restrictive mindset originates from an epistemic discredit of introspection, aligning with functionalism and behaviorism, which view inner life as a "hard problem" inaccessible to objective, scientific study. Outcomism can for instance lead to deny the very possibility of a distinction between genuine lived experience of compassion and the mere emission of compassionate behavior.

The danger is that as Generative AI excels at passing **Turing-style tests** (e.g., creating indistinguishable art, imitating moral experts), the outcomist focus rapidly shakes well-entrenched human intuitions, potentially discarding the legitimacy of accounting for the presence of human lived experience upstream of outcomes we are confronted with (factors like the presence of a human artist who elaborated a picture we contemplate).

To maintain robust ethical discernment, we must avoid this power grab. While reviewing basic beliefs is healthy, reducing the complex debate on (human) consciousness and agency to mere output comparison utilizes only a limited fraction of available (philosophical) arguments.

Taking lived experience, life and biology seriously

To avoid misleadingly blurring the fundamental distinction between humans and machines, we must challenge the core assumption that inner, lived experiences are kinds of "black holes" we can say nothing objective or reliable about (as "outcomism" does).

We can and must resist the systematic rejection of introspection. John Searle's "Chinese room" thought experiment, for instance, legitimizes the use of **introspective experience** to refute universal claims about the computational nature of the mind.

Furthermore, we must challenge the claim that phenomenal consciousness is solely a "hard problem" scientifically (objectively, seriously) approachable only through computational models that aims at reproducing the connections between inputs and outputs. Approaches like Antonio Damasio's demonstrate the possibility of enlarging the scope, studying consciousness not just through the brain's computational properties, but also through its grounding in the **subcomputational biological mechanisms** and the rich, organic interplay between the nervous system and the rest of the living body (notably via **interoception**, our inner perception of our own body).

By taking **life and biology seriously**, we can distinguish living beings from mere information technology artifacts. This perspective reinforces our basic intuitions regarding what is alive, conscious, and possesses autonomy or free will. These intuitions are far from old-fashioned obsolete prejudices that would be convincingly defused by serious scientific approaches. On the contrary, when not illegitimately reduced to outcomism, such scientific investigations rather confirm our intuitions can serve as reliable ground for further exploration.

Ensuring a robust understanding of humans' core specificities

We must critically address the common linguistic tendency to attribute human traits like **"intelligence," "decision-making,"** and a **"relationship to truth"** to AI and digital systems. While using these terms to describe automated functions may be convenient and admissible, it risks eroding the unique **ontological status** of humans, encouraging us to see ourselves as mere machines. Ethical governance requires that we **preserve the possibility** that these terms signify a deeper, non-reducible reality rooted in life and human lived experience.

Autonomy and Decision-making

Machine decision-making and autonomy, even in advanced AI, are valid only in a restricted sense. Computers are fundamentally **mechanical and inertial**, functioning as deterministic systems where response complexity is the only differentiator from a thermostat. They operate by rigid adherence to (possibly incredibly complex) algorithms and past data, always reacting the same way to the same input under a given state.

In stark contrast, **human autonomy** is an ontologically stronger concept, deeply rooted in **life and biological phenomena**. Human decision-making, which we know intimately through

introspection, transcends mechanical procedure. It is defined by the **ability to sidestep** past regularities and react differently to identical solicitations. This capacity for voluntary choice and **practical autonomy** is inseparable from our biological constitution, as well as from our lived phenomenal experience that includes our **affective and emotional life**.

This is paramount in the moral domain. Human moral decision-making is not mere re-application of past answers; it is a **power to make novelty.** This creative ability to take distance from established norms and genuinely consider new options is indispensable to acknowledge the essential possibility for a person to change and leave her past behind (a key component linked to human dignity). This openness and power to sidestep are core to decision-making in the strong, human sense.

Relationship to knowledge and truth

It seems undeniable that AI systems can produce true outcomes (most powerful systems are able to do so in ever growing ranges of topics and domains). Is this enough to attribute to them a relationship to knowledge and truth?

In the traditional philosophical understanding, knowledge is conceived of as **"justified true belief,"** which requires good reasons and justification beyond mere output of true statements. On this ground, some may claim that machines have a headstart over humans as they only apply logical-mathematical operations on raw data. They would thereby be endowed with a kind of perfect objectivity, a superior form of rationality freed by principle from any arbitrariness or subjectivity.

It is crucial to warn against such a distorted, though widespread view of rationality or intelligence as a kind of "mechanical objectivity"—a purely algorithmic process freed from subjectivity. In fact, history and philosophy of science reveal the limits of such approaches: the process of generating knowledge, even scientific, involves an **irreducible space of freedom** and the ineliminable activity of **informal judgment** by the knowing subject. There is no neutral, raw data; human judgments and arbitrations are indispensable for methodological choices and fundamental intuitions. Therefore, human intelligence involves not only applying criteria but **judging the quality of those criteria**.

Having a relationship to knowledge and truth in the strong human sense involves a critical, reflexive activity, which is fundamentally rooted in human **lived experience**. It is intimately tied to **autonomy** and **decision-making**, as it requires the ability to **sidestep** and imagine that admitted representations and beliefs could be different. Only this ability to sidestep makes humans sensitive to the call to make responsible use of their freedom and practical autonomy in a sincere quest for truth.

Some key topics for collective exploration

AI ethics and governance face a challenge deeper than merely mitigating power asymmetries between nations, between the public and the private sector, or between tech giants and users.

It is dangerous to assume the orientation for AI is obvious, as if the only issue were neutralizing malevolent actors.

A considerable, defining, task of AI regulation is the collective effort to define and articulate the **goals and objectives** AI should serve, rejecting the narrative that AI (or Super AI) is an unquestionable goal in itself. AI can be of service, notably in contributing to solve complex civilizational problems, but only if we first **define what we expect from it** and refine our human goals.

Therefore, our focus must shift from: **"What is our place as humans in the new world shaped by AI?"** to the essential question: **"What is the place of AI in the human world we want to build?"** This requires exploring many different topics. Among them, we would like to highlight and to encourage collective exploration of the particularly acute challenge of discerning how to position AI for it to preserve or even serve the flourishing of human core specificities.

Recommendation 8: Exploring how to assist and support humans in their relationship to knowledge and truth

Humans possess the core, fallible trait of relating to truth and collectively building knowledge—defined as **justified true belief**. Deep discernment efforts are required to determine how AI technology can genuinely **support** this process without undermining it.

For many components of information technology (such as online encyclopedia or journals, word processors or spreadsheets), we rely on strong **collaborative networks** where we delegate much of the quality assessment effort to trustworthy human experts (like journalists, encyclopedia editors or software developers) who warrant the reliability of the technology and of the results it presents to us. This cooperation leads, through division of labor, delegation and trust among humans, to a wonderful digital environment collecting and rendering accessible astonishing corpuses of knowledge (beliefs we have, direct or indirect, good reasons to hold true).

Generative AI presents a critical break in this respect. Since AI cannot be self-justificatory, only human autonomy can ultimately judge if reasons are good enough. And LLMs provide outputs that are **not warranted by a human's cognitive experience**. No specific subgroup checks the singular content delivered, shifting the entire burden of assessment onto the end-user.

Therefore, human communities need to develop **AI literacy** in relation to such issues associated with knowledge and truth. This literacy is necessary to distinguish among currently existing systems which can deliver **trustworthy pieces of knowledge** (warranted by human experts) from those that are merely powerful tools for exploration. It is also indispensable for exploring the type of new systems we may develop to bring additional dedicated support in the various facets of our cognitive and epistemic lives. Strong reflection is required to assign AI its proper place for preserving, fostering and prolonging our collective efforts to relate to truth.

Recommendation 9: Reflecting upon the manner AI can serve human (collective) intelligence

To ensure AI genuinely serves humanity, we must look beyond the mere reliability of tools and address the preservation and development of **human intelligence**. This involves a two-pronged approach regarding individual skills and our collective environment.

First, we must reflect upon the risks of **deskilling**. While delegating tasks to AI can be efficient, overuse can impede cognitive development. A thorough analysis is required to identify which skills and lived experiences are indispensable—not just for verifying AI outputs, but for maintaining the broader cognitive faculties necessary for human flourishing.

In addition, AI has become a **"cognitive extension"** of the human mind. Predictive algorithms and generative AI contribute to the editorialization of our reality, shaping the "informational substrate" upon which we think. Because human knowledge is fundamentally **collective**—relying on a "common decency" (the will to judge and know in common, committing to the validity of beliefs before others)—the way AI structures this environment is critical.

Currently, the contribution of AI to this collective intellectual life is worrying. Driven by an economic model based on **attention capture**, AI systems often generate **cognitive bubbles** and **echo chambers**. This causes "epistemic harm," degrading the trust and benevolence required for a healthy collective intellectual life.

This toxicity is not a fatality. AI holds tremendous potential to broaden perspectives and foster mutual understanding and common decency. However, realizing this potential requires a deep **political and collective effort**. We must move beyond "fixing" or regulating current algorithms from the outside and fundamentally **re-orient economic models** away from attention capture, designing AI to serve as a fertile ground for genuine human collective intelligence.

Recommendation 10: Exploring how AI can contribute to human agency and responsibility

The rise of complex AI systems, particularly **Agentic AI** capable of executing actions, necessitates a proactive focus on ensuring **meaningful human control**—a concept broader than simply solving legal **responsibility gaps**. The real problem lies in avoiding a **control gap** where powerful, unpredictable mechanical systems become misaligned with human values and objectives. This requires robust **AI literacy**, enabling communities to discern which tools are trustworthy for automation and which (like Generative AI) demand strict supervision to prevent dangerous, unpredictable outcomes.

A profound ethical threat to this agency is the **temptation of cognitive offloading**. Fueled by "Promethean shame" (a sense of inferiority to machines) and an aversion to risk, humans may delegate decisions to AI to avoid the burden of responsibility and the possibility of error. This surrender is incompatible with **decision-making in the strong human sense**, which relies on the capacity to **sidestep** past regularities, exercise creativity, and imagine alternative

possibilities. Maybe more threatening than an abrupt takeover, we face the risk of a **"gradual disempowerment"**—an incremental erosion of human influence over societal systems.

To combat this, collective effort is required to resist the temptation to reduce oneself to an inertial object illusorily relieved from any responsibility or painful exercise of autonomy. We must design systems that **empower rather than replace** human judgment. We must discuss key questions: **What price are we willing to pay to defend human autonomy?** How do we ensure algorithms support and empower our strong decision-making abilities, rather than encouraging the illegitimate offloading of responsibility?

Recommendation 11: Problematizing the notions of progress, good life and vulnerability

To ensure AI truly serves the "good life," we must resist **techno-solutionist shortcuts** that uncritically link technological innovation with genuine human progress, assuming every problem has a technological fix. This mindset risks reducing human issues to metrics of efficiency, ignoring root causes (such as the societal sources of loneliness or professional burnout) in favor of superficial technological patches. Such shortcuts, rooted in a kind of **"technocratic paradigm,"** reduce reality to manipulable indicators, viewing human progress solely in terms of **efficiency** and measurable performance.

This mechanistic mindset is problematic because it mutilates the legitimate search for human freedom by treating any limit, mistake, or **vulnerability** as a defect to be eliminated. This pursuit of infallibility and unlimited power is epistemologically and morally flawed.

As previously established, strong human **knowledge and decision-making** are intrinsically fallible, requiring an essential **margin of maneuver** and the possibility of making mistakes. The freedom to choose comes with the risk of choosing wrong. "Improving" humans by removing this fallibility through automation does not enhance human agency and relationship to truth; it eradicates them. True progress lies in refining our responsibility and critical thinking, not in an illusory quest for infallibility.

More fundamentally, it is crucial to acknowledge the subtly of the notion of vulnerability. Vulnerability cannot be reduced only to its negative aspects (‘vulneration’ like injury or illness). Vulnerability also corresponds to the fundamental **possibility of being affected**. While we have a duty to prevent injury, trying to eliminate vulnerability is a mistake. Absolute robustness is the negation of life; it is our vulnerability that allows us to love, feel joy, and connect with others.

Therefore, we should turn away from the misleading question about whether AI can make us all-powerful, infallible, and invulnerable. Instead, the collective reflection must ask: **How can we develop and use AI systems to help us better tame and balance the ambivalent but essential vulnerability and fallibility that lies at the deepest heart of who we are?**

Recommendation 12: Cultivating our sensitivity to life and conscious lived experience

To correctly guide AI development and use, we must **cultivate our sensitivity to life and conscious lived experience**, actively assessing when the **presence of a genuine, vulnerable person** is indispensable.

Such discernment necessitates strong collective effort and can prove extremely challenging. It notably demands resisting the philosophy of **outcomism** threatens this effort by focusing exclusively on results and external behavior. Outcomism can lead to the problematic justification of using AI for social roles by equating convincing **appearance** for authentic **presence** (LLM-generated medical communications could be sufficient as they possess apparent empathy, AI-generated art is legitimate if it can pose for human generated one, AI companions could help coping with loneliness even better than humans as they would never abandon or oppose their users …).

Resisting such reductionist approaches is indispensable and perfectly legitimate. As already exposed, genuine human presence is indispensable for knowledge elaboration and decision-making, with the irreducible need for responsible use of freedom. It seems primary in contexts like healthcare and psychotherapy, where shared affectability is central to connection and effective therapeutic outcomes. One could mention even more straightforward examples: a child's imperfect drawing, representing hours of effort and intention, holds more value than a flawless image generated instantly by an AI. Words of compassion expressed to a dying person have no value if they do not signal a genuine lived experience of compassion.

Thus, when striving to discern where to deploy AI systems and for what usages, it is perfectly legitimate (and absolutely indispensable) not only to account for the 'objective' quality of outcomes, but also to deeply reflect upon the possible value of the **presence of a genuine, vulnerable, and affectable person** in the elaboration processes upstream. This implies that people must be informed when interacting with systems that convincingly mimic humans (transparency and disclosure of AI use), as this knowledge is key to enabling reflection on the value of genuine presence.

Ensuring that AI is put at the genuine service of human relationships and sensitivity to life will thus impose confronting with ambiguous and subtle cases. The case of AI companions illustrates well the depth of difficulties. It seems clear that they cannot replace genuine relationships. The latter necessitates the presence of a genuine human person with her autonomy, which gives all the value to interaction, despite (or rather because of) the risk of friction and abandonment such freedom implies. However, one may argue that, used as mere toys (like imaginary characters in books, films or videos games we might get attached to), AI companions become innocuous. But things may be more subtle because of the level of human imitation these systems can reach, which triggers the risk of **schooling users in the negation of the other** and fostering a culture that views intimacy as a schematized commodity, potentially making users less tolerant of the true autonomy and friction inherent in genuine human relationships. In the same vein, proposing to deploy AI systems to offload overburdened professionals (in the healthcare context for instance) might be a way to reduce professionals' exhaustion, but it could also become a means of avoiding confronting the root causes of encountered difficulties.

In sum, collective reflection to orientate AI development and use must confront with these important questions about how to foster and cultivate, in the age of AI and, if possible, with the support of AI, our sensitivity to life and to the presence of lived experience of genuine persons. More than that, we must cultivate our **ability to assess the decisive value of genuine presence over mere appearance**, and to discern where and how it is primary, and how it should be balanced with considerations about outcomes' quality.

Detailed analysis

Introduction

When it comes to AI, saying that the last few years have been frenetic would be quite an understatement. From the technical standpoint, the rise of deep machine learning, more recently reinforced by the transformers technology, led to many wonderful achievements. Among them, Large-Language Models (LLM) and generative AI have spread throughout society at an unprecedented pace and sparked as much hope as anxiety. Extremely enthusiastic actors do not hesitate to announce general, or even super, AI (AI systems that reach or even exceed human levels in any cognitive task) for the coming decade(s), joined in their prophetic stance by more alarmist ones who warn against the existential risk this would pose for humanity. In a far more tangible and actual fashion, AI is largely presented as one of the main economic drivers of the current era. One must learn how to surf on the coming wave. Companies must implement AI as deeply as possible in all their activities to maintain their competitivity. To remain (or become) prominent, nations and tech actors must race and develop as fast as possible the most powerful LLMs and generative AI systems. Any attempt at even slightly constraining regulation generates its share of protests, portraying the regulatory effort as economic suicide. As the French sociologist Dominique Boullier claims, our societies got trapped in a 'cognitive tunnel' marked by a kind of '**tyranny of tardiness**'.[1]

Such a hectic societal atmosphere does not constitute an ideal framework for ethical development and use of AI, and all the more so since it happens on the ground of **growing power asymmetries** in favor of tech giants and associated financial circles. Recent studies documented the manner some major American tech actors and investors gained more and more influence over democratic and governance processes, notably because of their privileged position in the editorialization of our informational landscapes but also because of their involvement as subcontractors or solution providers in national public domains such as security.[2] Some of these actors either promote or adhere to techno-solutionist ideologies, as strikingly exemplified by the venture capitalist Marc Andreessen's Techno-Optimist Manifesto ('We believe that there is no material problem – whether created by nature or by technology – that cannot be solved with more technology').[3] As Hans Jonas long put it, such uncritical

[1] Dominique Boullier, 'Sommet IA : la nécessaire sécession sémantique européenne - AOC media', *AOC media - Analyse Opinion Critique*, 9 February 2025 <https://aoc.media/analyse/2025/02/09/sommet-ia-la-necessaire-secession-semantique-europeenne/> [accessed 21 October 2025].

[2] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2020); Rob Lalka, *The Venture Alchemists: How Big Tech Turned Profits into Power* (Columbia University Press, 2024); Marietje Schaake, *The Tech Coup: How to Save Democracy from Silicon Valley* (Princeton university press, 2024).

[3] Marc Andreessen, 'The Techno-Optimist Manifesto', *Andreessen Horowitz*, 16 October 2023 <https://a16z.com/the-techno-optimist-manifesto/> [accessed 24 June 2025].

relationship to technology may allow for technological development to 'continuously [gather] new momentum, carrying its carriers along as its appointed instruments.'[4]

In this context, democratic governance of technology development in general and of AI in particular appears more indispensable than ever (despite its challenging nature, especially with respect to economic imperatives). This need is well perceived by the public opinion.[5] As Zuboff stated summarizing the spirit she sees in the European Digital Services Act, '**the digital must live in democracy's house**.'[6] This being said, it is important to make clear that the challenge of democratic regulation does not reduce to power imbalances. Difficulties also stem from the pace and uncertainties of AI development and its consequences, generating for instance the so called 'evidence dilemma' with the need to balance between excessive 'pre-emptive risk mitigation measures based on limited evidence' and the danger of 'waiting for stronger evidence of impending risk.'[7] Moreover, the modalities of regulation raise multiple questions: what equilibrium between legally binding instruments and soft law tools? What principles and values should guide the regulation? What about citizens' and stakeholders' involvement? Under which forms? How to ensure enforcement of enacted regulations?

More than a challenge of mere regulation, what our societies are confronted with is the urge to commit to a 'long overdue work of reinvention.'[8] During its first three years of operation (2022-2025), the members of the NHNAI network put their resources at the service of such a collective effort of exploration, especially through the prism of the topic of humanism and the question of what it means to be human in the age of AI. Drawing on the findings and learning of this first phase of operation, this white paper intends to propose some recommendations to approach the challenge of fostering development and uses of AI systems that would be at the genuine service of humanity. These recommendations will be organized according to three core axes: 1) recommendations on the manner the background reflection aimed at supporting regulation should be conceived of and organized, 2) recommendations on basic contents that could constitute the ground of this background reflection (elements without which the reflection could be impaired), 3) recommendations on important topics that should be explored through this background reflection.

What background effort to support strong democratic regulation?

[4] Hans Jonas, *Philosophical Essays: From Ancient Creed to Technological Man* (Prentice-Hall, 1974), p. 48.

[5] 'There is a strong public mandate for AI regulation, with 70% believing regulation is necessary. However, only 43% believe current laws are adequate. People expect international laws (76%), national government regulation (69%), and co-regulation with industry (71%). 87% also want laws and fact-checking to combat AI-generated misinformation.' Nicole Gillespie and others, *Trust, Attitudes and Use of Artificial Intelligence: A Global Study 2025* (The University of Melbourne and KPMG, 2025), p. 5, doi:10.26188/28822919.

[6] Shoshana Zuboff, 'Democracy Can Still End Big Tech's Dominance Over Our Lives', TIME, 5 May 2022 <https://time.com/6173639/democracy-big-techs-dominance-shoshana-zuboff/> [accessed 25 October 2025].

[7] Yoshua Bengio, *International AI Safety Report* (2025), p. 14.

[8] Zuboff, 'Democracy Can Still End Big Tech's Dominance Over Our Lives'.

It is now largely acknowledged that AI regulation needs inputs from societal actors. The important question then becomes determining what type of efforts and reflections should be settled to produce these inputs. It is particularly important to discuss who should be involved and to do what, with what purposes.

An inclusive and horizontal reflection (recommendation 1)

With respect to the first question, it is important to point out the limitations of any technocratic or elitist settings. One could for instance be tempted to delegate the effort of regulation design only to groups of experts (from scientific and technical domains as well as from human and social sciences, …) and political representatives who would adopt a position of epistemic surplomb. However, nothing warrants that such groups, though indispensable, can alone arrive at correct answers. It seems important to deploy more inclusive efforts, mobilizing all concerned stakeholders. Independently of this need for stakeholders' involvement with respect to the quality of the content of regulation themselves, too weak social participation would lead to low social awareness on the very existence and content of regulatory devices while it is a key component of robust enforcement of regulations.[9]

A reflection not only on uses of AI systems but also on their design (recommendation 2)

An even more important issue lies in the aims, purposes and modalities of this participation for designing societal inputs to regulate AI. Though already well known, a point deserves to be recalled here: **governance of technology (and societal inputs thereby required) does not reduce to regulating uses**. A bit more in-depth discussion might be useful as this claim, as obvious it may sound on the surface, conflicts with our cultural tradition (marked by the nature-culture dichotomy and the tendency to picture ethics as an individual deliberation activity on one's own choices and actions). In fact, one must resist the traditional view (partly inherited from the modernity) according to which, on the one hand, technological development and the scientific activity it relies on pertain to the domain of *facts*, neutrally producing truths on what *is* and means for action; and on the other hand, that society (individual and democratic will) reigns over the realm of *values*, freely deciding about what *should be* done with these knowledge and powers. According to this fact-value dichotomy, societal (individual or democratic) will merely complements the picture by choosing what to do with neutral facts and technological means. However, this view is far too simplistic. Many critics attacked the very possibility of a fact-value dichotomy.[10] Ensuring strong democratic and societal regulation of AI necessitates to properly understand the subtle relationships and **entanglements between technology and ethics** (understood as the reflection, deliberation and action in the field of what *ought to be*).

[9] A recent worldwide study from the University of Melbourne and KPMG demonstrated that '[m]ost people are unaware of laws, legislation or government policy that apply to AI.' See: Gillespie and others, *Trust, Attitudes and Use of Artificial Intelligence*, p. 8.

[10] See for instance: Hilary Putnam, *The Collapse of the Fact-Value Dichotomy, and Other Essays Including the Rosenthal Lectures* (Harvard University Press, 2002).

As the quite long-standing techno-critic tradition made clear, artifacts almost always embed political and ethical implications.[11] Thus, most of the time **ethics begins well upstream of usage**, right from the design stage of technologies. This is especially true of AI technologies, which often have greater autonomy compared to more traditional artifacts. Concerns with the possible loss of control over powerful generative AI systems constitutes an extreme illustration,[12] picturing disasters without any use choice per se, just because the technology exists. As striking and worrisome such perspectives may be, we should not let them obfuscate the fact that the need for ethical reflection upstream applies in a far more widespread way.

As Moor proposed in 2006,[13] we should consider most digital devices as 'ethical-impact agents', to the extent they produce outcomes of ethical relevance. This manner of framing the issue makes clear how far upstream ethics can go. Ethical reflection is somehow there the very moment one tries to build a *good* artifact. In a sense, the very objectives of reliability and security of our computers and programs are already ethical stakes engineers are tasked to cope with through artifacts design. Features ensuring varying degrees of privacy correspond to in design ethical choices we are more used to. Returning to AI, one could mention the well-studied cases of biased classification tools or algorithmic decision systems[14] (especially when they are mobilized for the administration of public services), or of recommendation algorithms that editorialize the huge amount of information available on our liberalized digital information markets (on internet or social platforms) according to attention capture purposes.[15] Here again, ethical 'quality' of outcomes does not depend only (or even mainly) on choices at the level of uses. **Ethical quality is largely built in the design of systems**.[16]

Based on this conceptually structuring reminder, it is therefore clear that regulation and ethical reflection must also happen at the early stage of AI systems development and not only at the

---

[11] Langdon Winner, 'Do Artifacts Have Politics?', *Daedalus*, 109.1 (1980), pp. 121–36.

[12] A recent open letter, which received the support of many public personalities and famous researchers, expresses deep concerns associated with the possible emergence of super AI, 'ranging from human economic obsolescence and disempowerment, losses of freedom, civil liberties, dignity, and control, to national security risks and even potential human extinction,' see: 'Statement on Superintelligence', Statement on Superintelligence, 2025 <https://superintelligence-statement.org> [accessed 26 October 2025]. For a less mediatic but more thorough analysis, see: Bengio, *International AI Safety Report*, sec. 2.2.3.

[13] J.H. Moor, 'The Nature, Importance, and Difficulty of Machine Ethics', *IEEE Intelligent Systems*, 21.4 (2006), pp. 18–21, doi:10.1109/MIS.2006.80.

[14] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, First paperback edition (B/D/W/Y Broadway Books, 2017); David Restrepo Amariles, 'Algorithmic Decision Systems: Automation and Machine Learning in the Public Administration', in *The Cambridge Handbook of the Law of Algorithms*, ed. by Woodrow Barfield, Cambridge Law Handbooks (Cambridge University Press, 2020), pp. 273–300, doi:10.1017/9781108680844.015.

[15] Gérald Bronner, *Apocalypse cognitive* (PUF, 2021).

[16] Moor proposes a useful additional distinction. When outcomes of ethical relevance are obtained or generated directly because of the system design, Moor talks about 'implicit ethical agents'. This contrasts with 'explicit ethical agents' who explicitly manipulate ethical elements (as could be moderation algorithms or tools to debias databases that would for instance rely on explicit ethical rules and criteria). This second category of systems corresponds to a full-fledged sub-field of AI called 'machine ethics'. See for instance: *Machine Ethics*, ed. by Michael Anderson and Susan Leigh Anderson, 1st ed. (Cambridge University Press, 2011), doi:10.1017/CBO9780511978036.

level of uses of ready-made solutions. As Dignum phrases it, 'Responsible AI' necessitates **ethics for, in and by design**.[17] And again, citizen's and stakeholders' participation is necessary as inputs for external regulation of AI systems development processes but also within these development processes, through dedicated of R&D such as Value Sensitive Design (VSD), Responsible Research and Innovation (RRI) or Participatory Design (PD).[18]

A transformative effort of genuine political reflection, with the support of experts (recommendation 3)

In fact, there is an ineliminable **political dimension of the design of AI systems**, a dimension that must be acknowledged in all its thickness. Especially, societal inputs and stakeholders or citizen participation cannot reduce to the mere neutral collection of people's mindsets through opinion polls or sociological surveys. One can hardly assume they will always provide actionable inputs, ready and legitimate for guiding design processes. First, being sure to include all legitimate voices and represent what they say properly can prove extremely difficult. Moreover, preferences, values and ethical principles people adhere to may enter in conflict (the question of surveillance assisted by AI facial recognition illustrates well such tensions, with the tension between the goals of improving security and of privacy and freedom preservation). Although extremely valuable,[19] mere snapshots of public opinions can be nothing more than starting points.

In many of the deep ethical issues AI raises, the manner society is at the given instant, preferences people have, are more part of the problem than of the solution. Biased classification algorithms are so largely because of the examples society provided in the first place (we return to this point in more length in the next section). In the same vein, the threat that recommendation algorithms pose to our ability to make good use of our free attention time is to a large extent caused by the free economic model (uncritically adopted by many customers). Thus, solving such issues thus not only means properly guiding AI design and uses, but also necessitates social transformations for rendering possible this guidance. In addition,

[17] Ethics in, by and for design respectively correspond to the implicit and explicit layers discussed by Moor, and to the ethical context surrounding the design work (economic pressure, gender balance in software engineering teams for instance). See: Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Artificial Intelligence: Foundations, Theory, and Algorithms (Springer International Publishing, 2019), doi:10.1007/978-3-030-30371-6.

[18] For instance, the UE AI HLEG recommends mobilizing technical as well as non-technical methods with 'Stakeholder participation and social dialogue' to build 'Trustworthy AI throughout the system's entire life cycle'; see: High-Level Expert Group on AI, *Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future* (2019), pp. 20–23 <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [accessed 8 August 2025].(p. 20-23). For a more academic discussion and details upon participatory methods, see: Dignum, *Responsible Artificial Intelligence*; Till Winkler and Sarah Spiekermann, 'Twenty Years of Value Sensitive Design: A Review of Methodological Practices in VSD Projects', *Ethics and Information Technology*, 23.1 (2021), pp. 17–21, doi:10.1007/s10676-018-9476-2; Carolyn Ten Holter, 'Participatory Design: Lessons and Directions for Responsible Research and Innovation', *Journal of Responsible Innovation*, 9.2 (2022), pp. 275–90.

[19] Such as with the already mention report from The University of Melbourne and KPMG that 'aims to provide an evidence-based understanding of people's trust, use and attitudes toward AI, their views on the impacts of AI, and expectations of its governance and regulation' (Gillespie and others, *Trust, Attitudes and Use of Artificial Intelligence*, p. 4.

these transformative efforts will often require expert inputs from specialists of many different disciplines (philosophers, economists, psychologists, historians, computer scientists …) to support participants in their reflection. In this respect, and as the examples above illustrate well, many AI ethical issues can be seen as '**wicked problems**', complex societal problems 'such as violence, hunger, poverty, disease, and environmental pollution', that are as much technical and scientific (including the human sciences) as they are political.[20] They constitute deep political and ethical interrogations and challenges about how societies should be organized and the manner individual persons should live and at the same time embed technical and scientific questions.

Accordingly, elaborating societal inputs susceptible to guide regulation and design of AI often demands 'transdisciplinary' co-production of solutions,[21] where specialists from many different disciplines put their expertise at the service of collective reflections in which citizens and stakeholders are the central actors. In sum, participation must include, with the assistance of experts, processes of **self-criticism, mutual enrichment and transformation of social groups themselves**.[22] Again, this should be distinguished from any technocratic approaches where experts, from a dominant epistemic position, would teach societal actors what to do. Experts only bring some stones for the edifice the collective must build. They support reflexivity, self-criticism and self-transformation of the collective and benefit themselves from these (as researchers, but also as citizens).

Saying, as we just did, that social groups must enter into reflexive and transformative processes to build *better* input for AI regulation and development may sound quite problematic. It enters in conflict with the cultural tradition we mentioned above. Indeed, the fact-value dichotomy is often invoked to justify the idea that – contrarily to factual issues that can be investigated empirically or scientifically, and thereby settled in a compelling way – ethical, political, and more broadly, evaluative questions do not pertain to the realm of knowledge, rationality and truth, but to the one of pure freedom and free-will. Therefore, individuals are free to adopt whatever view they want about what *should be*. And while anybody else is totally free to disagree, nobody can tell someone else he or is wrong (tell them their claims are *false*). We agree to disagree. We just don't have the same values. At most, we can try to rally someone else to our interests for purely pragmatic reasons (in a sophist way), but not in the logic of a collective quest for truth, for improved shared views (precisely what we have been trying to defend the need for). Here again the fact-value dichotomy is both wrong and dangerously misleading. First, philosophy of science and epistemology of the second half of the 20th century made quite clear that scientific and factual investigations irreducibly mobilize evaluative

[20] Christian Pohl, Bernhard Truffer, and Gertrude Hirsch-Hadorn, 'Addressing Wicked Problems through Transdisciplinary Research', in *The Oxford Handbook of Interdisciplinarity*, ed. by Robert Frodeman (Oxford University Press, 2017), p. 0, doi:10.1093/oxfordhb/9780198733522.013.26.

[21] Julie Thompson Klein, 'Typologies of Interdisciplinarity: The Boundary Work of Definition', in *The Oxford Handbook of Interdisciplinarity*, ed. by Robert Frodeman, 2nd ed. (Oxford University Press, 2017), pp. 21–34 (sec. 3.5), doi:10.1093/oxfordhb/9780198733522.013.3.

[22] Florin Popa, Mathieu Guillermin, and Tom Dedeurwaerdere, 'A Pragmatist Approach to Transdisciplinarity in Sustainability Research: From Complex Systems Theory to Reflexive Science', *Futures*, 65 (2015), pp. 45–56, doi:10.1016/j.futures.2014.02.002.

judgments.[23] Moreover, it is (obviously?) far from obvious that there cannot be evaluative (including ethical and moral) knowledge or rationality.[24] Therefore, there is legitimacy to talk about ***better* societal inputs in the strong sense** (not just better relatively to some particular groups or interests).

Of course, our intention here is not to contest the legitimacy of democratic freedom and democratic pluralism. What we believe should be opposed is the tendency to caricature them into relativistic or solipsistic views that decouple this freedom from the responsibility and the duty to sincerely seek truth or validity in everything we think, claim and do. As Revault d'Allonnes explains well, this decoupling is at the root of post-truth abuses and, while it may give the illusion of enhanced freedom, it in fact undermines people's ability to truly inhabit their world (as they become unable to recognize factual evidences about the manner the world is, which is a precondition to imagine other ways it could be and begin transforming it).[25] Applied to our topics of the elaboration of societal inputs to guide regulation, development and use of AI, we cannot rest content (in the name of abstract democratic pluralism) with and are not condemn to stop at a mere patchwork of more or less diverging preconceived opinions (each actors simply tolerating conflicting opinions and for instance abiding to the will of the majority). We have the possibility, and therefore the duty, to settle **communities of rational discussion and deliberation**[26] working at elaborating ethical understanding and guidance with respect to AI societal issues. In particular, instead of systematically treating divergent claims as matters of individual free opinions, we may attempt at generating some basic agreement on the fact that at least some divergences are the sign of legitimate **tensions or complexities inherent to the question being explored**. Take for instance the goals of security and privacy or fundamental right protection that conflict when it comes to AI powered surveillance. Although people may differ about which one to prioritize, it seems possible for all to agree on the legitimacy of both objectives. Deploying efforts in common to rationally establish such solid ground would undoubtedly improve the collective ethical reflection.

Ethical capacity-building for bottom-up and horizontal guiding forces (recommendation 4)

In a sense, the reflexive and transformational efforts needed to produce societal inputs susceptible to guiding AI development and regulation can be seen as ethical capacity-building processes. By deploying sincere and collective work to map the ethical issues and their complexities and to try to build answers, participating communities will cultivate their capabilities to contribute to ethical guidance of AI. Such empowerment is indispensable for multiple reasons. First it is needed for communities of citizens and stakeholders to help

[23] Julian Reiss and Jan Sprenger, 'Scientific Objectivity', in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta, Winter 2020 (Metaphysics Research Lab, Stanford University, 2020) <https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/> [accessed 27 May 2025].

[24] Aaron Zimmerman, *Moral Epistemology*, New Problems of Philosophy (Routledge, 2010), doi:10.4324/9780203850862.

[25] Myriam Revault d'Allonnes, *La faiblesse du vrai: ce que la post-vérité fait à notre monde commun*, La couleur des idées (Éditions du Seuil, 2018).

[26] Philip Kitcher, *Science, Truth and Democracy*, Oxford Studies in the Philosophy of Science (Oxford university press, 2001).

designing top bottom general ethical principles and legal regulations. But beyond that global dimension of regulation, the presence of communities with reinforced ethical capacities are also crucial for contextualizing and concretely applying general ethical or legal principes in local contexts. It can considerably facilitate the application and enforcement of legal regulations (local ascending awareness and expectations meeting descending regulatory efforts). Local communities can also contribute to render more efficient soft law devices and similar types of more horizontal non legally binding regulation devices (labels, charters, guidelines, …). Enlightened consumer choices will also be key, opening viable economic spaces for more ethical entrepreneurship.

In addition, and maybe importantly, empowered communities can participate in the identification of needs and in the definition of high-added values uses cases of available technology. We will develop some aspects of this question a bit below, but we must already highlight here the danger there would be in believing in a kind of principled usefulness of AI technologies (especially of the prophesized general or super artificial intelligence). Various wonderful technological building blocks are available, and no doubt new ones will be developed. However, finding truly positive use cases of them might be one of the most serious challenges of AI ethics and regulation. Enlarged and enlightened contribution from all concerned stakeholders at this crucial level of AI development will be decisive to get AI technologies truly contributing to human societies in their various dimensions. In summary, ethical capacity-building as widespread as possible is necessary to foster regulation and ethical development and use of AI by creating or reinforcing **top-down, as well as bottom-up and horizontal guiding forces**.

A dedicated exploration on what being human means (recommendation 5)

Let's now turn toward the content, the topics on which these ethical capacity-building efforts may bear. In this respect, it is particularly important to frame the general questions in the right causal order. Too often we hear injunctions for humans to find their place in the new world shaped by AI. Joshua Bengio and the other co-authors of the *International AI Safety Report* (2025) put it very clearly (and we will return to this in the next section): 'AI does not happen to us: choices made by people determine its future'.[27] This means that the good question is "what is the place of AI in our human world our natural environment?". In the ideal, we should give orientation to AI by deeply reflecting upon our needs, which in turn implies global effort to build shared understanding concerning the society projects AI should serve.

In the following, we would like to focus on an important dimension of these questions. In fact, to answer them, one must also reflect upon what "being humans" means, on what we are, but above all on what we want or need to be as humans. This convokes the theme of humanism in its various senses (not only descriptive, historical or cultural, but also axiological). Numerous major principles of AI ethics appeal to the idea of the human: AI must be *human*-centric, at the service of *human* flourishing, the *human* must be kept on or in the loop, etc. More fundamentally, we can even see an almost organic link between ethics and the quest to understand what it means to be human. Ricoeur, for example, defines 'ethical aim' as 'the aim

[27] Bengio, *International AI Safety Report*, p. 14.

of the "good life" with and for others in just institutions'.[28] How can we begin to forge the meaning of terms like "good life" or "just institutions" without at the same time reflecting on what it means to be human? Moreover, new technologies such as AI (but we could include also neurotechnology in the picture) provide humanities with increased powers of in-depth self-modification (an *anthropotechnè* within the framework of which humans can influence their own nature).[29] How should we transform ourselves, in which direction? A reliable compass about humanism and what it means to be human seems more necessary than ever.

And it is indeed an effort to explore and deepen that is at hand. The notion of humanism is far from being clear and consensual, unproblematic and ready to serve the purpose of ethical orientation. Even taken only in the context of its emergence (in Europe with the Renaissance, then Modernity and the Enlightenment), the notion already incorporates multiple, disparate and sometimes conflicting dimensions.[30] Subsequently, many currents have opposed and continue to oppose the humanism of modernity head-on (anti-, post-, trans-humanism), highlighting its difficulties and limitations. Are there precise characteristics that distinguish the human from the non-human? Are they universal? Isn't the almost absolute primacy granted by modern humanism to human autonomy and rationality problematic? Hasn't it led to the myth of the human as master and possessor of nature, with an automatic link between techno-scientific development and human progress? Today, these difficulties seem to be reinforced by the theoretical and scientific contributions of AI and NS, as well as by their technological spin-offs.

But perhaps the solution is not to reject the notion of humanism outright. It is also possible to put the notion back to work, through a renewed exploration capable of preserving and deepening the most fruitful contributions of the humanism of Modernity and the Enlightenment. It is common, for example, to retain from Kant only the idea of a human reason that can reach *a priori* conclusions in the spheres of science (about phenomena) and morality (with the categorical imperative) - pure theoretical and practical reason. It is therefore all too common to point to the failure of Kantian epistemology (for example, the overcoming of Newtonian physics) and to discredit Modernity and humanism. But this would be to ignore the heart of Kantianism, with its central idea of the passage of humanity from minority to majority (the individual can and must think for himself, "*Sapere Aude*") and the importance of the fundamental couple freedom-responsibility, which leads Kant to place practical reason, and

[28] Paul Ricoeur, *Soi-même comme un autre*, L'Ordre philosophique (Ed. du Seuil, 1990), p. 202, our translation: 'visée éthique' is defined as 'la visée de la "vie bonne" avec et pour autrui dans des institutions justes'.

[29] Sylvain Lavelle, 'What a Human Is, Could Be and Should Be. The Anthropology of the Human and the Philosophy of Humanism', in *Human Freedom at the Test of AI and Neuroscience*, ed. by Stefano Biancu, Mathieu Guillermin, and Fabio Macioce, Contemporary Humanism: Open Access Annals (2024) (Edizioni Studium, 2024), pp. 119–41 <https://www.edizionistudium.it/riviste/studium-contemporary-humanism-open-access-annals-2024>.

[30] See for instance the illuminating historical and philosophical presentation of the notion of humanism: Tony Davies, *Humanism* (Routledge, Taylor & Francis Group, 2001).

above all the faculty of judgment, at the center of his entire philosophical system.[31] A « critical » rather than dogmatic modernity is conceivable.[32]

Depriving ourselves of this type of input would only fuel the difficulties with AI ethics, governance and regulation. It seems far more fruitful to deepen these resources, and to couple them with the exciting insights of AI, neuroscience and cognitive science in order to outline the contours of a new humanism, opening up to a renewed understanding of our freedom, our intelligence, or our capacity to judge. To support the exploration of this question of what it means to be human, it may first be useful to chart some fundamental milestones about the reality of AI technology and thereby about the possibility of pointing out certain specificities of humans by comparison with machines.

Fundamental milestones on AI and humans

To explore, build and deliberate upon society projects AI should serve, based on a thorough collective exploration of what it (should) mean(s) to be human, a precondition is to ensure a robust and empowering understanding of what AI is. Unfortunately, as evoked in the introduction, AI is often misrepresented or pictured in a too abstract way. It is for instance common to present various successes machine (deep) learning permitted to produce as successes of **a unique big thing called AI**. In particular, it is very impressive to picture that big AI as, on the one hand, capable of producing new (and possibly better) manners of achieving a task (AlphaGo and the discovery of a new powerful way to play Go for instance) and, on the other hand, able to answer all our questions, especially difficult ones (such as with recent versions of GPT that can answer PhD level questions in natural sciences).

This way of setting the stage sends the message that, with enough data and computing power, AI has basically no limits. On this type of ground, many announce the foreseeable emergence of Artificial General Intelligence (AGI) or even of Super Artificial Intelligence (SAI), fueling the idea of a (ultimately lost) competition between humans and AI in the field of intelligence. Background beliefs in these matters are often quasi-religious, as illustrated by Eric Schmidt recent statements: SAI is so because it 'can prove something that we know to be true, but we cannot understand the proof'.[33] This type of atmosphere also nurtures the tendency to frame the problem of the possible loss of control over AI systems in terms of a conflict or war between intelligences. Even Geoffrey Hinton, the recipient of the 2024 Nobel prize in Physics (for fundamental research that paved the way to modern machine learning techniques), claimed the problem of control is critical as 'there are very few examples of more intelligent things being controlled by less intelligent things' (one of the rare examples being a mother being controlled by

[31] Alexis Philonenko, *L'œuvre de Kant: la philosophie critique. 1: La philosophie pré-critique et la critique de la raison pure*, A la recherche de la vérité, 6. ed (Vrin, 1996); Alexis Philonenko, *L'œuvre de Kant: la philosophie critique. 2: Morale et politique*, A la recherche de la vérité, 5. ed (Vrin, 1997).

[32] Bernard Feltz, *La science et le vivant: philosophie des sciences et modernité critique*, 2e éd. revue et augmentée (De Boeck, 2014).

[33] Eric Schmidt is former CEO of Google. See: Eric Schmidt, 'AI and the Genesis of a New Epoch', Public conference, RAISE Summit 2025, Paris, 8 July 2025 <https://www.youtube.com/watch?v=_gBxYL2ihc0> [accessed 30 October 2025].

her baby).[34] Such representations of **AI as a kind of inexorable wave that will render humans cognitively obsolete** are particularly harmful, especially as **they completely screen off the political and ethical dimension of AI as a human artefact**. Their toxicity is reinforced by the often low level of literacy lay people have about current AI technologies.[35] In consequence, it is essential to secure a not too abstract understanding of AI within communities engaged in the effort of producing resources and guidance for its development and use.

Ensuring a robust and empowering understanding of AI technologies (recommendation 6)

Here are some contributions to a more empowering representation of AI.

Demystifying machine learning

Machine learning is a technical notion and field that is often oversimplified for the broad audience, for instance by saying that it is about 'baby' machines or programs that will somehow learn to perform a task as a child who learns to do something new. In addition, machine learning is sometimes presented as producing black boxes whose functioning humans cannot understand. And as we just said, these opaque systems would learn to answer all our questions, especially in finding better solutions to many of our problems. This type of representation is extremely problematic. **Much more can and should be said about machine learning**. One must at least explain that **machine learning amounts to automatically searching for an adequate parametrization of an computational architecture**, in the hope that we may end up with a program capable of performing a task that was resisting to explicit programming. For instance, it is rather simple to write an algorithm for classifying simples images, say monochromes of different colors (comparing the average of the values encoding colors in each pixel of each image would do the trick). However, we do not know what operations the computer should do to correctly classify images of multiple ordinary objects. In the case of such resisting tasks, **machine learning** techniques may allow to **partially bypass our programing limitations**.

In fact, maybe we cannot prescribe step by step the operations to do to perform a classification of images of real-world objects. But what we can do select some classes of operations (for instance multiplications by some coefficients, additions and other mathematical operations on numbers specifying colors in the pixels of an image) and bet that there are arrangements of these operations that would do the job. This amounts to designing a **computational architecture with *free parameters*** (values of multiplying coefficients for instance). This

[34] Geoffrey Hinton, 'Will Digital Intelligence Replace Biological Intelligence?', Public conference, Romanes Lecture, University of Oxford, 20 February 2024 <https://www.ox.ac.uk/news/2024-02-20-romanes-lecture-godfather-ai-speaks-about-risks-artificial-intelligence> [accessed 30 October 2025].

[35] As reported in a recent study from a journalist consortium, many people (here 30% to nearly 50% depending on the age) tend to trust AI as a reliable source of information, which it is not. Even more worrisome, more than a third of respondent tend, when receiving a false information because the LLM failed at summarizing properly, to blame not only the AI, but also the original news source. See : European Broadcasting Union (EBU), *News Integrity in AI assistants* (2025), p. 4 <https://www.ebu.ch/fr/research/open/report/news-integrity-in-ai-assistants> [accessed 24 October 2025].See also : Mark Steyvers and others, 'What Large Language Models Know and What People Think They Know', *Nature Machine Intelligence*, 7.2 (2025), pp. 221–31, doi:10.1038/s42256-024-00976-7.

architecture becomes many different programs when the details of the arrangement of selected operations are determined by setting the free parameters to specific values. **The bet with machine learning is then the following**: we assume (hope based on an educated guess) that at least one of these programs we can get from a specific parametrization of the architecture we designed can properly perform the task we want. Then we can write a more or less smart trial and-error program whose task will be to test various sets of parameters to fine the most efficient one (or at least one efficient enough). When this is achived, we say we *learned* a model or a program. And now maybe the most important: **all of this can work only by providing a guidance or a feedback** to this parameters tuning program. For machine learning to even be possible in the principle, one must design a way of quatifying the quality of the outcomes produced by a given set of parameters.

This manner of explaining machine learning techniques has the merit to make clear why the idea of a unique big powerful thing called AI is illusory. First, there are very **different types of computational architecture** we can try to use and *train* that are more or less well suited to specific problems (convolutional neural networks, reccurent neural networks, transformers, ... to name only but a few). Moreover, **the manner the feedback that guides the learning process is defined can also greatly vary and open the way for different types of successes and thereby of expectations**. In the case of board games such as Go (as well as with most of video games), the feedback is easy to design: one can just use the score at the end of a match or a game. In such cases, the automatic parametrization may well lead to a program that finds novel ways of maximizing the feedback (new powerful ways to play Go humans were not aware of), thereby possibly producing (notably) wonderful tools for assisted exploration. But these are niches configuration wherein the feedback can be directly and explicitely framed in algorithmic or mathematical terms (even if in some case the algorithm is very big, as in the case of simulators or video games). However, in most of real life scenarios, such as for LLM and conversational bots, but also for picture classification, we do not have such an directly algorithmic or mathematical definition of the feedback. What we have instead are sets of examples (conversations, books, already classified pictures ...) that we can use as a golden standard the learning process tries to reproduce through various parametrization of the computational architecture. But, in such cases (and they are the most widespread), maximizing the feedback is obtained by reproducing the examples. So it is pointless to expect (and misleading to claim) that it will produce radically new results humans were themselves unable to produce before (on the contrary, a parametrization producing new results would not be selected, being poorly rated as it does not reproduces the examples). At best, this type of machine learning based on examples can leat to programs with a limited power of generalization, for instance through the generic ability one might get from the capacity to hold credible conversations.

In addition, the deabstracted view of machine learning we proposed allows shedding some light upon the 'fallacy of inscrutability', consisting in presenting (deep) machine learning as

producing 'mysterious, unaccountable black-box software systems'[36]. First, it should now be clear that some humans (engineers or developpers) know pretty well the type of operations the software is doing (for instance, a multitude of multiplications, additions, and not-so-complex other mathematical functions in the case of deep neural networks). This, however, does not mean that these programs are free from any opacity. For most of deep learning algorithms, it is for instance very difficult to understand, by looking at the program itself, whether and why it will work. There are far too many operations, and they work at the sub-symbolic level (on list of numbers that represent manipulated objects in high-dimensional spaces). The so-called "hallucinations" of LLMs and the possibilities for prompt injections or other adversarial attacks[37] illustrate quite well the type of 'surprises' such an opacity can hide. And these 'surprises' deep learning algorithms can produce are very difficult to predict and anticipate *a priori*. Contrary to most of our tools (including more analytically engineered computer programs, such as with symbolic AI), we lack in this case a reliable theory of error.[38] In that sense, there can be a real and serious issue of control with some deep learning algorithms.

This is nevertheless not a reason to just ban these algorithms. On the one hand, a very active sub-field of AI is focused on this topic of explainability[39] and progress seems possible for interpreting in an informative way what we see of the functioning of deep neural network programs such as LLMs. On the other hand, nothing (technical) prevents an empirical *in situ* testing of such programs. This aspect is key for an enlightened use of deep learning programs, in particular as it allows for distinguishing between two broad configurations. While it is reasonable to hope to gather enough examples of inputs-outputs in the case of specialized programs, such as tools for classifying medical images for a specific pathology, things may become trickier when the task becomes more complex (classifying images of any possible objects) or even fuzzy (answering correctly to any question). The more expected results vary, the more difficult it will become to ensure that empirical testing covers enough ground to be reliable. This means that reliability cannot always be expected (especially not in the case of generative AI and LLMs) and uses must be adapted in consequence. In cases where reliability cannot be warranted enough, we should refrain from unsupervised delegation of tasks as well as from misled interfacing of unreliable and unpredictable programs to the rest of our information systems (as sometimes incautiously promoted with the growing trend of agentic AI). This is typically the type of dangers Bostrom's 'paperclips' thought experiment permits to

---

[36] Joshua A. Kroll, 'The Fallacy of Inscrutability', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376.2133 (2018), p. 20180084, doi:10.1098/rsta.2018.0084.

[37] Yujie Sun and others, 'AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content', *Humanities and Social Sciences Communications*, 11.1 (2024), p. 1278, doi:10.1057/s41599-024-03811-x; Xiaoyong Yuan and others, 'Adversarial Examples: Attacks and Defenses for Deep Learning', *IEEE Transactions on Neural Networks and Learning Systems*, 30.9 (2019), pp. 2805–24, doi:10.1109/TNNLS.2018.2886017.

[38] Daniel Andler, *Intelligence artificielle, intelligence humaine : la double énigme*, NRF Essais (Gallimard, 2023), sec. 4.5, Cairn.info, doi:10.3917/gall.andle.2023.01.

[39] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda, 'Transcoders Find Interpretable LLM Feature Circuits', *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA), NIPS '24, 37 (2024), pp. 24375–410.

highlight. Such scenarios are often convoked to illustrate the threat of AI becoming smarter than humans. Not only the possibility, but also the very meaningfulness of the idea can be debated (see recommendation 7). There is, however, a crucial element paperclips-like scenarios put in plain sight: the mechanical and inertial nature of AI.

Recalling the materiality of AI

In fact, one of the main problems with paperclips like scenarios is that the machine cannot deviate from its program if something goes wrong. But this is not a software issue. On the contrary, such an inertia is inherent to the hardware and to the very idea of computation. Too often, we picture AI and digital technologies as largely immaterial (with terms such as 'cloud', 'dematerialization' and abstract representations involving colored lines of code, data or mathematical equations). However, we should never forget that **all programs (from the most traditional and conventional to the most advanced AI program produced by machine learning) run on computers or similar machines that are not (or less) programmable**. For such automatic computation to be possible, humans must first establish some conventions that associate meanings to material traces or configurations (for instance a series of magnets on a hard drive disk whose orientations symbolize a sequence of 0s and 1s, itself associated, for example, with a sequence of words or a sequence of numbers coding the colors of pixels in an image). Then, very capable humans can design machines like computers that will transform in a precisely controlled way (reflecting an algorithm or a program) these material traces into new ones associated with other meanings (for example, a new series of words, a modified image or a description of the image). Presented that way, it becomes obvious that one of the first properties we expect from such machines is to never deviate from its intended functioning as prescribed by the program. This type of machines, designed to transform material configurations into others according to what these configurations signify, is not new. The computer can be seen as the culmination of a long history of information techniques and technologies, probably dating back to the very beginnings of writing. From this perspective, the abacus can be seen as an ancestor of the computer (mechanical transformation of configurations symbolizing, for example, numbers to be added, into configurations symbolizing the result of addition).

So, strictly speaking, there are no meanings, images, words or numbers in computers, let alone emotions or consciousness. They are, however, fantastic machines for mechanically manipulating (with incredible efficiency and precision) countless material configurations to which we humans attach meaning. A series of magnets on a computer hard drive disk will cause different pixels on the screen to emit different colors, which will be more than just tiny sources of colored light for us, which will become texts telling us about feelings, images of faces feeling such and such emotions... But the computer only processes information by mechanically and automatically manipulating magnets (or other hardware configurations). This makes it all the more breathtaking to see what we can get computers to do with programs derived from machine learning techniques.

Highlighting the direct dependence on human intelligence

It is perfectly legitimate to marvel at AI (as well as to worry about it). However, the sense of wonder we may feel with AI technology must be for the right reasons. As we just saw, its

successes have nothing to do with the creation of new forms of life, new intelligent beings that would compete with us, we would call the AIs. It is just as dizzying, if not more so, to realize that mankind has been able to build machines, artifacts capable of simulating or reproducing intelligent behavior (convincing behavior that could have come from humans), with absolutely no life, no lived experience, no consciousness, but with pure mechanisms (inert mechanisms, but dazzlingly complex and miniaturized). The impressive successes of AI should never make us forget that AI is a product of human intelligence, directly dependent on it.

The exposition proposed up to now makes clear, we hope, that the notion of Super Artificial Intelligence (as coined by Eric Schmidt, a system providing outcomes we know to be true without being able to understand the proof) is largely illusory. In the current state of AI technology, there is no magical warranty that produced outcomes are valid. Machine learning techniques do not change the picture. Humans remain in charge of building computer programs and of assessing their results. It takes a very high level of human intelligence to get good AI systems. Of course, it necessitates very smart programmers, engineers and computer scientists for designing computational architectures and learning procedures. But, designing feedback for the guidance of learning process also demands very smart people, and in vastly broader communities than the programming work itself. Formal framing of this feedback can be very tricky, calling for a lot of knowledge from people who are experts of the targeted domain. Building reliable databases of examples for indirect reconstruction of feedback mobilizes the intelligence of numerous persons (journalists, contributors to Wikipedia, literature authors, researchers publishing books or articles in academic journals, …, anybody producing clever content on digital platforms and information systems).

Irreducible human responsibility also remains at the level of AI systems quality and reliability assessment. We need a lot of human efforts and intelligence to build trust in a given digital technology. Very often laypeople delegate this assessment work, but not to machines themselves. We delegate it to specialists we trust. And in the case of generative AI, it is well-established that the level of reliability is too low to blind trust outcomes (as illustrate very well the Terms of use of these systems that are crystal clear about the user responsibility over generated results). Returning to the control problem, it can be very difficult to warrant that very complex and powerful generative systems won't answer favorably to illegitimate or illegal demands (such as support for criminal action or dangerous advice to psychologically fragile people). It is also a very difficult problem to give good objectives to learning processes and good prompts to very complex and powerful systems in order to ensure they won't produce dangerous unexpected results. In addition, it may require a lot of human strength and intelligence to refrain from using, in not secured enough settings, systems that would not present sufficient warranties of reliability. In a more global way, giving adequate orientation to AI and warranting AI systems comply with our goals presupposes humans have been clever enough to define these orientations and goals (as discussed in recommendations 1 to 5). This may be the place where the highest level of collective human intelligence is needed. As we shall turn to now, such a crucial endeavor may also benefit from some anthropological and philosophical insights.

Securing some basic intuitions on the specificities of humans by comparison with machines (recommendation 7)

We developed the point in recommendations 1 to 5, and previous section made its importance even clearer. What AI will become, the services or traps and threats it will present to us are direct consequences of the quality of the human reflection that will preside over its development and uses. This reflection aiming at a refined discernment on the place of AI in our societies and daily lives requires not only a demystified grasp of AI, but also a robust approach of what it means to be human, what is core to humanity, of our strengths and our limits. In particular, we need a sound and shared understanding of our specificities as humans by comparison with AI systems.

In this respect, some experts defend very disruptive and counter-intuitive claims. In recent interviews, Yan Le Cun or Stanisla Dehaene indicated they see no principled opposition to machines becoming conscious.[40] Recent scientific reports similarly claimed that the possibility of AI consciousness cannot be excluded, and that one should thus consider seriously the topic of AI welfare.[41] Should we, as a growing number of voices urge us to, reconsider our basic intuitions about the ontological status of AI systems and computers? Should we start envisaging that machines may be endowed with core human traits such as consciousness, intentionality, or free will? Although blind dogmatism should be resisted (here as everywhere), revising basic intuitions and beliefs of this importance should be done only with great caution, for extremely good reasons. It is far from obvious that this is presently the case.

Resisting the injunction to "outcomism"

Many of the rationales leading to question common intuitions and to claim AI may reach a new ontological status take their roots in and get traction from a focus only on outcomes and results humans and machines can produce. In fact, the line of reasoning relies on the idea that we can have direct contact with inner experience (phenomenal consciousness, will, intentionality, …) only through introspection, which is not an objective and scientifically reliable source of knowledge. As Chalmers put it, the phenomenal aspects of mental life constitute the 'hard problem' or the 'hard part' of the mind-body problem, in contrast to psychological properties and behavioral dimensions that can be studied objectively, from the outside, through a functionalist approach: with 'functional properties characterized by causal roles, so the question "How could a physical system have psychological property P?" comes to the same thing as "How could a state of a physical system play such-and-such a causal role?"'[42] This is

[40] Le Cun (*Le Point*, 2023, https://www.lepoint.fr/sciences-nature/intelligence-artificielle-le-debat-choc-et-inedit-harari-le-cun-11-05-2023-2519779_1924.php); Dehaene said that "consciousness is a computational property", *Le Point*, 2023, https://www.lepoint.fr/sciences-nature/intelligence-artificielle-a-quand-une-conscience-artificielle-30-08-2023-2533358_1924.php).

[41] Patrick Butlin and others, 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness', arXiv:2308.08708, preprint, arXiv, 22 August 2023, doi:10.48550/arXiv.2308.08708; Robert Long and others, 'Taking AI Welfare Seriously', arXiv:2411.00986, preprint, arXiv, 4 November 2024, doi:10.48550/arXiv.2411.00986.

[42] David John Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Philosophy of Mind Series (Oxford university press, 1996), sec. introduction and I.1.4.

such a defiance with respect to the scientific reliability of introspection that led Turing to propose the famous 'imitation game' as the only legitimate test for assessing whether machines could think.[43] For behaviorism and reductionist forms of functionalism inner or mental states can be approached only as causal connections between observable phenomena (that can include brain states described as internal states of computational machines). As Janet Levin aptly puts it, approaches of this type 'do not threaten to denote, or otherwise induce commitment to, properties or processes (directly) observable only by introspection'.[44]

This epistemic discredit cast upon introspection paves the way to what we can call "outcomism," prescribing to restrict the discussions about the possibility for machines to possess specific human traits pertaining to the phenomenal domain (consciousness, intentionality, will, …) to the comparison between outcomes AI systems and humans can respectively produce. Danaher's ethical behaviorism illustrates well the restriction: we should not ask AI to prove its phenomenal consciousness in a more strict and demanding way we do for a friend or another fellow human.[45] Such an outcomist approach prohibits, on principle, any distinction between the emission of words and behaviors expressing compassion and the expression of a genuine compassionate experience. This outcomist approach is very potent in shaking most well-entrenched intuitions. Generative AI systems become increasingly capable at passing artistic Turing tests (humans becoming unable to determine whether productions are generated by AI or not).[46] A study demonstrated a tendency to perceive intentions and expression of emotions pieces of art people know to be generated by AI.[47] The outcomist mindset can even push some authors to deny the legitimacy of taking into account the knowledge of who made a piece of art when judging its quality (they see this as an anthropocentric prejudice and claim that we should 'debias people's perceptions of AI art', to allow 'accessing things at their true form (…) free from prejudice and preconceptions').[48] The same line of argumentation can also be made about the moral domain. LLMs seem able to imitate even renowned moral experts such as The Ethicist from the New York Times.[49]

---

[43] Alan M. Turing, 'Computing Machinery and Intelligence', *Mind*, LIX.236 (1950), pp. 433–60, doi:10.1093/mind/LIX.236.433.

[44] Janet Levin, 'Functionalism', in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta and Uri Nodelman, Summer 2023 (Metaphysics Research Lab, Stanford University, 2023) <https://plato.stanford.edu/archives/sum2023/entries/functionalism/> [accessed 28 November 2025].

[45] AI Research Group of the Centre for Digital Culture, 'Encountering Artificial Intelligence: Ethical and Anthropological Investigations', *Journal of Moral Theology*, 1.Theological Investigations of AI (2023), pp. 74–80, doi:10.55476/001c.91230.

[46] Brian Porter and Edouard Machery, 'AI-Generated Poetry Is Indistinguishable from Human-Written Poetry and Is Rated More Favorably', *Scientific Reports*, 14.1 (2024), p. 26133, doi:10.1038/s41598-024-76900-1.

[47] Theresa Rahel Demmer and others, 'Does an Emotional Connection to Art Really Require a Human Artist? Emotion and Intentionality Responses to AI- versus Human-Created Art and Impact on Aesthetic Experience', *Computers in Human Behavior*, 148 (2023), p. 107875, doi:10.1016/j.chb.2023.107875.

[48] Kobe Millet and others, 'Defending Humankind: Anthropocentric Bias in the Appreciation of AI Art', *Computers in Human Behavior*, 143 (2023), p. 107707, doi:10.1016/j.chb.2023.107707.

[49] Danica Dillion and others, 'AI Language Model Rivals Expert Ethicist in Perceived Moral Expertise', *Scientific Reports*, 15.1 (2025), p. 4084, doi:10.1038/s41598-025-86510-0.

Functionalist frameworks can be employed to defend the attribution of moral agency to some AI systems.[50]

While it is true that there are intense debates about the profound nature of phenomenal consciousness, intentionality and free will and about their possible attributions to machines, outcomist positions of the type just discussed are far from consensual. They constitute a very limited portion of broader and very dense discussions.[51] No doubt these debates are relevant for ethical discernment and should not be considered as purely philosophical.[52] But it would be misleading to mobilized only a reduced fraction of their argumentative dimensions. We must resist such power grab. Again, it's cognitively healthy to be ready to revise basic beliefs and intuitions. But we should do so only in the presence of strong reasons. It is far from obvious that the mere existence and possibility of philosophical positions negating differences between humans and machines in these matters makes it legitimate to discard well-entrenched intuitions we have. Moreover, the outcomist strategies discussed here rely on the assumption that phenomenal consciousness, inner lived experiences and associated features are epistemic blackholes that cannot be studied seriously in themselves. This very assumption itself may not be as solid as it looks at first sight.

Taking lived experience, life and biology seriously

First, we can point out the precipitous nature of a systematic rejection of introspective elements. Searle's argumentation in his famous "Chinese room" thought experiment is particularly interesting in this epistemological perspective.[53] This text is often presented as a more or less successful attack against functionalism. This is not the aspect we are interested in for the present discussion. Rather, what matters here is the place that Searle gives to introspective experience in his argument, which can be reconstructed as follows: the thesis that "to have a mind is to execute the right kind of program" is aimed at *all* minds (universal quantification). From this, we can logically *deduce* that the thesis applies to any singular mind (mine, the reader's, Searle's). The Chinese room thought experiment allows us to introspectively convince ourselves that our own mind doesn't work that way. Each person doing this

---

[50] Morgan S. Porter, 'Moral Agency in Silico: Exploring Free Will in Large Language Models', arXiv:2410.23310, preprint, arXiv, 28 October 2024, doi:10.48550/arXiv.2410.23310.

[51] See for instance: John R. Searle, 'Minds, Brains, and Programs', *Behavioral and Brain Sciences*, 3.3 (1980), pp. 417–24, doi:10.1017/S0140525X00005756; Hubert Lederer Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (MIT Press, 1992); Hubert L. Dreyfus, 'Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian', *Artificial Intelligence*, 171.18 (2007), pp. 1137–60, doi:10.1016/j.artint.2007.10.012; Adrien Doerig, Aaron Schurger, and Michael H. Herzog, 'Hard Criteria for Empirical Theories of Consciousness', *Cognitive Neuroscience*, 12.2 (2021), pp. 41–62, doi:10.1080/17588928.2020.1772214; Sara Lumbreras and Eduardo Garrido-Merchán, 'Insights from Saint Teresa and Saint Augustine on Artificial Intelligence: Discussing Human Interiority', *Scientia et Fides*, 12.2 (2024), pp. 265–95, art. 2, doi:10.12775/SetF.2024.025; Andrzej Porębski and Jakub Figura, 'There Is No Such Thing as Conscious Artificial Intelligence', *Humanities and Social Sciences Communications*, 12.1 (2025), p. 1647, doi:10.1057/s41599-025-05868-8.

[52] Mario De Caro, 'Does Imputability Require Free Will? The Discussion in the Civil Law Tradition', in *Human Freedom at the Test of AI and Neuroscience*, ed. by Stefano Biancu, Mathieu Guillermin, and Fabio Macioce, Contemporary Humanism: Open Access Annals (2024) (Edizioni Studium, 2024), pp. 41–53 <https://www.edizionistudium.it/riviste/studium-contemporary-humanism-open-access-annals-2024>.

[53] Searle, 'Minds, Brains, and Programs'.

introspective experiment gets directly acquainted with a counterexample refuting the universal thesis that bears upon *all minds*. Given the nature of the thesis, a refutation through introspective experience is perfectly legitimate.

Although it may constitute a significant theoretical element for resisting against outcomism, this case in favor of the epistemic legitimacy of introspection does not suffice to provide a firm ground to discuss the reliability of our intuitions concerning the specificities humans keep by comparison to machines when it comes to phenomenal consciousness, intentions or free will. Fortunately, more can be said. Notably, one can question the claim that phenomenal consciousness and related mental features cannot be studied directly from the outside, but only through their causal effects or reduced to a computational only model of the brain. A different picture may emerge when taking into account 'subcomputational biological mechanisms'.[54] This is precisely the line Damasio explores.[55] For him, phenomenal consciousness constitutes a hard problem only because it is approached with a too strong focus on the brain and its computational properties or descriptions. Once we accept to enlarge the scope, it becomes possible to describe phenomenal consciousness as grounded in the very basic activity of living organisms (homeostatic processes constituting a kind of 'non-explicit intelligence') and as emerging because of the rich and organic interweament between the central nervous system and the rest of the living body. A key element for Damasio is interoception that, far from a mere mental representation of what happens in the body, results from a dynamic dialogue between neurons and the rest of the living tissues and organs. Interoception constitutes the basic texture of the self-conscious mind, a background of feeling within which other mental experiences will happen.

Despites their possible limits, such approaches evidence the possibility and the legitimacy to take life seriously, at least through the lens of biology. In this perspective, computers and related machines belonging to the domain of information technology can be straightforwardly distinguished from living beings. The latter are specific just in virtue of being living organisms. There seems to be no compelling reasons to abandon our basic intuitions about what is alive or not (at least when it comes to computers). In the same vein, our common intuitions on who is consciousness and what is not seem reliable enough. When we recognize the legitimacy of biology and branches of neuroscience still interested in biology, scientific investigations may lead to refine them but in no case refine them, especially when it comes to artifacts from information technology. The same type of discussion can be conducted for the topic of free will. For sure, we are far from perfectly understanding the notion. However, it seems important to make room for biological inputs. Living organisms with central nervous systems seem to possess a kind of autonomy and ability to sidestep inert computational automata are deprived from.[56]

---

[54] Ned Block, 'Can Only Meat Machines Be Conscious?', *Trends in Cognitive Sciences*, published online 7 October 2025, doi:10.1016/j.tics.2025.08.009.

[55] Antonio R. Damasio, *Feeling & Knowing: Making Minds Conscious*, First Vintage Books edition (Vintage Books, 2022).

[56] Björn Brembs, 'Towards a Scientific Concept of Free Will as a Biological Trait: Spontaneous Actions and Decision-Making in Invertebrates', *Proceedings of the Royal Society B: Biological Sciences*, 278.1707 (2010), pp. 930–39, doi:10.1098/rspb.2010.2325.

Ensuring a robust understanding of humans' core specificities

In light of the previous sections, it seems important to discuss in more depth the widespread linguistic practices that tend to attribute to AI and digital technologies traits usually associated to humans (and possibly some animals). In fact, it is quite common to talk about automated *decision-making* or about LLMs revolutionizing the *relationship to knowledge* or *to truth*. The term "artificial intelligence" itself suggests the idea that we build *intelligent* machines. While it may be legitimate to talk in such a way according to some senses or definitions of these terms, it would be a dangerous mistake to reduce them to these limited conceptions. Not only would this allow claiming that machines can reach (part of) the ontological status of living beings and of humans. Even more worryingly, it would tend to reduce the ontological status of living beings and humans to the one of computers and AI systems. It would encourage seeing living beings and humans as nothing more than mere machines. Nevertheless, we just saw that nothing prevents taking life and biology seriously, as well as lived experience we can scrutinize through introspection. If nothing prevents it, maybe it may be a duty to do so. It seems possible, and thereby necessary for ethical and political discernment, to clearly affirm some core specificities of human beings, at least by comparison to AI systems and computers.

Autonomy and Decision-making

In this perspective, we should always remember that talking about automated decision-making or AI autonomy can only be valid in a very limited sense. Computers and other information technology artifacts are purely mechanical and inertial machines. So, what we mean by machine autonomy can be nothing more than a more or less complex reaction to variations in inputs they receive. Of course, some programs can update the internal memory of a computing machine that may then react differently to further input. Some programs can even modify other programs or select which program to run in given circumstances. But all of these, seen globally, remain deterministic and mechanistic responses. A computer in a given internal state always reacts the same way to the same input (at least that's what we expect from computers, what we build them for … and we deploy a lot of effort to compensate when they malfunction in this respect, for instance with Error-Correcting Codes). In terms of autonomy, computers with or without AI programs belong to the same realm as thermostats. They differ only in the degree of complexity of their responses.

When we talk about autonomy and decision-making for living beings and humans, we mean more than that, something stronger, ontologically different. Especially when considering humans (whose autonomy and decision-making process each of us is acquainted to from within through introspection), it seems obvious enough that being autonomous or deciding is not just about applying algorithms or procedures in a mechanical way. Endowed with their characteristic kind of autonomy, living beings can react differently to same solicitations. Conscious living beings like humans can do that with practical autonomy and free will. At least, this is what decision-making means in the strong sense (a sense we just saw nothing compels to abandon or to deny to humans): voluntarily *choosing* between available options. Humans have the ability to decide about what *should be* when confronted with a plurality of possibilities. And as Damasio famously demonstrated it with the 'somatic markers' hypothesis, this ability

to decide in the strong sense does not rely on computational features our brains may have but irreducibly involves affects and emotions.[57]

What we just developed also applies to moral decision-making. Producing (as LLMs seems capable of) sentences that convey moral advice or judgements is not the same as genuinly possessing moral expertise and being capable of moral decision-making in the strong sense. And the question to determine whether the moral content of sentences produced by LLMs is valid or not misses the point. The correctness of this content is relevant for other topic (see below the discussion of the role of artificial moral advisors) but is orthogonal to the present reflection. When humans decide in the moral domain (as in any other domains) they do more than merely re-applying past answers to moral problems. They mobilize their ability to sidestep, to genuinely consider options. They do their best to choose the best one. They may also sense that something is wrong in the past ways of doing or past norms. As Dominique Lambert puts it, humans are capable of 'creativity' in the sense of a 'power to make novelty'.[58] This power to sidestep, to take distance with past regularities is key from the moral point of view. Machine learning techniques can lead to AI systems aptly predicting what people may do based on what they did in the past. But, as Pope Françis recalled with strenght, '[a]lgorithms must not be allowed (…) to eliminate the possibility of an individual changing and leaving his or her past behind.'[59] Only humans are able to maintain open and deal with such possibilities. Such a creative power is core to (moral) decision-making in the strong sense.

Relationship to knowledge and truth

We can now turn to the topic of knowledge and relationship to truth. How to understand claims about the fact that AI revolutionizes our access to knowledge? Is it meaningful to think AI could produce better knowledge than us? Should we really revise our basic intuitions on knowledge and relationship to truth being core specific traits of human beings? We already provided some deflating elements on this question based on the reality of machine learning techniques (see recommendation 6). But more can be said when looking at what it means, in the strong sense for humans to know and to have a relationship to truth. Here, as before, it is crucial to resist any outcomist approach. Knowledge is not just a set of true statements. Producing knowledge is not reducible to elaborating true claims. Traditional definitions are straightforward on this: 'knowledge is *justified* true belief'.[60] Similarly, 'a person, *S*, knows that *p* (where *p* is a proposition) if and only if (i) *S* believes that *p*, (ii) *S* has justification (evidence, good reasons) for *p*, and (iii) *p* is true.'[61] Knowledge necessitates justification, good reasons to believe a given claim.

---

[57] Antonio R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (Penguin, 2005).

[58] Our translation. Dominique Lambert, 'Retrouver l'humain Au Cœur de l'IA et de La Robotique, et Lui Redonner Toute Sa Place: Conférence', *Revue Confluence*, N° 6.2 (2024), pp. 23–42, doi:10.3917/confl.006.0023.

[59] Pope Françis, *Artificial Intelligence and Peace*, Message for the 57th World Day of Peace (1 January 2024), https://www.vatican.va/content/francesco/en/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html

[60] James Ladyman, *Understanding Philosophy of Science* (Routledge, 2002), p. 6.

[61] Gürol Irzik, 'Critical Rationalism', in *The Routledge Companion to Philosophy of Science*, ed. by Martin Curd and Stathis Psillos, Routledge Philosophy Companions, Second ed (Routledge, 2014), pp. 70–78.

As recalled in recommendation 6, current LLM technology cannot be blind trusted. LLMs cannot be said to produce knowledge themselves as humans always have to check the validity of outputs themselves. However, SAI prophets do not claim the superintelligence is already here. They just argue that it is possible and will happen in the foreseeable future. It may be that such predictions get some traction from the idea that computers, precisely because they are not alive, have a principled head start over humans. As they only apply logical-mathematical operations on raw data, computational machines would be endowed with a kind of perfect objectivity, a superior form of rationality freed by principle from any arbitrariness or subjectivity. Arrived at this point, it is important to say that such a line of reasoning relies on a very specific (though widespread) conception of intelligence and rationality: to be rational or objective is to purge investigation procedures of any contingent content, any elements that could be different and would necessitate making a choice and thus evaluating options. In such a conception of rationality as 'pure enquiry',[62] any specificity of subjects must be removed. Knowledge and truth must be pursued through a kind of 'mechanical objectivity' exclusively based on empirical measurement and algorithmic or logico-mathematical procedures.[63] Grounded in this type of approach, it's indeed tempting to imagine that machines based on ever increasing computing powers and amounts of data could at some point reach a superior form of relationship to knowledge and truth.

However, recent history and philosophy of science (since at least the second half of the 20th century) has shown us the limits of such a purely algorithmic or procedural conception of rationality and intelligence. The notion of justification (of good reason to consider a belief as a knowledge) is not entirely amenable to mechanical objectivity. Any effort for elaborating some knowledge, even the most scientific and experimental ones, inevitably encompasses an irreducible space of freedom and implies an ineliminable activity of informal judgment from the behalf of knowing subjects. There is no such things as raw data and neutral logical-mathematical procedures that would impose themselves. Human judgments and arbitrations are indispensable (for instance concerning the basic vocabulary to be used, the major methodological orientations, the objectives to be achieved... but also concerning fundamental intuitions such as the idea that empirical observation does not systematically deceive us).[64] As Hilary Putnam puts it, with Cavell, knowledge, and more broadly 'speaking and thinking

[62] Bernard Arthur Owen Williams, *Descartes: The Project of Pure Enquiry*, [Rev. ed.] (Routledge, 2005).

[63] Reiss and Sprenger, 'Scientific Objectivity', sec. 4.

[64] Kitcher, *Science, Truth and Democracy*, p. See for instance:; Ernan McMullin, 'The Virtues of a Good Theory', in *The Routledge Companion to Philosophy of Science*, ed. by Martin Curd and Stathis Psyllos, Routledge Philosophy Companions, Second ed (Routledge, 2014), pp. 561–72; Mathieu Guillermin, 'Non-neutralité sans relativisme? Le rôle de la rationalité évaluative', in *Et si la recherche scientifique ne pouvait pas être neutre?*, ed. by Laurence Brière, Mélissa Lieutenant-Gosselin, and Florence Piron (Éditions Science et bien commun, 2019), pp. 315–38; Pierre-Luc Dostie Proulx and Mathieu Guillermin, 'The Role of Explanatory Virtues in Abduction and IBE', in *Logic in Question*, ed. by Jean-Yves Béziau and others, Studies in Universal Logic (Springer International Publishing, 2022), pp. 471–90, doi:10.1007/978-3-030-94452-0_24; Kyle Stanford, 'Underdetermination of Scientific Theory', in *The Stanford Encyclopedia of Philosophy*, ed. by Edward N. Zalta and Uri Nodelman, Summer 2023 (Metaphysics Research Lab, Stanford University, 2023) <https://plato.stanford.edu/archives/sum2023/entries/scientific-underdetermination/> [accessed 2 June 2025]; Catherine Allamel-Raffin, 'Modes de raisonnement et images scientifiques', *Visible*, no. 11 (May 2023), doi:10.25965/visible.165.

subjects', cannot exist without contextualized and situated acts of 'acknowledgement' and 'attunement'.[65]

Therefore, to be intelligent or rational, to entertain a relationship to knowledge and to truth, is of course to be able to correctly (objectively or neutrally) apply criteria, procedures or algorithms, but it is also, and perhaps above all, to be able to judge the quality of criteria and procedures, to have a reflexive and critical attitude towards what we are doing... and therefore to be able to judge and arbitrate fallibly, to make mistakes sometimes, to correct oneself, to evolve (and to help each other in this respect, to collaborate with good will)... Being intelligent in this strong sense is something fundamentally alive, something that each of us can only undertake rooted in our own lived experience (with all the richness but also the limits that this entails)[66] and in healthy collaboration with others. On top of that, what we just exposed shed some light on the intimate connection between having a relationship to knowledge and truth in the strong sense and being capable of decision-making (also in the strong sense). Knowing in the human sense irreducibly involves being confronted with available options among which one must choose. It implies practical autonomy within an essential space of freedom. Humans can have a relationship to knowledge and truth because of their ability to sidestep, to conceive things (here their representations and admitted beliefs) could be different from what they are. Only this ability makes humans in position of knowing in the strong sense that involves jugding as best as possible, without absolute certainty, which among the available options looks the most reasonable. Only this ability to sidestep makes humans sensitive to the call to make responsible use of their freedom and practical autonomy in a sincere quest for truth.

Some key topics for collective exploration

In light of the content discussed up to this point, we can refine our understanding of the challenge of AI regulation. It would be dangerous to reduce this challenge to a question of power asymmetries between countries, between tech giants and users, etc., as if the orientation we should give to AI was obvious and the problem was only to neutralize malevolent actors who have strong interests in pushing in other directions. The assumption that we know where AI should go is more than debatable. It is in fact a considerable part of the challenge of AI regulation to define the goals and objectives AI should serve. Some may try to propagate the narrative according to which AI (general or super AI) is a goal in itself as, once reached, it would have the potential to solve all our problems. However, it should now be clear enough that this sort of magical AI is a complete fantasy in the current and foreseeable state of AI technology. No doubt machine learning techniques and AI systems can solve many problems and help mitigate the most acute civilizational issues we are confronted with. They already do. But AI

[65] Hilary Putnam, 'Philosophy as the Education of Grownups : Stanley Cavell and Skepticism', in *Philosophy in an Age of Science: Physics, Mathematics, and Skepticism*, ed. by David MacArthur and Mario De Caro (Harvard university press, 2012), pp. 552–64.

[66] François Laplantine, *The Life of the Senses: Introduction to a Modal Anthropology*, Sensory Studies Series (Bloomsbury Academic, 2015).

can be of service that way only if we are able to define what we expect from it, to refine our goals and the manner we want to reach them. And this is far from always straightforward.

Issues we face are often very complex and sometimes an overly hasty response with AI can be ill-adapted (or even reinforce existing problems). We could think of the problem of the lack of personnel in retirement homes. Of course, some well-designed robots may help care workers to save time and go faster. But should we not also reflect upon the cause of the understaffing? Is the work recognized enough? Paid enough? Developing and deploying AI solutions should always be guided by thick design processes involving concerned stakeholders to refining the understanding of the singular problem at hand and what could mean a genuine solution or mitigation for it. These processes should also pay careful attention to choosing the right piece of technology and adapting it to the singular purpose at hand. This in particular means not trying to put generative AI or deep learning tools without discernment. Sometimes they can be inappropriate. Sometimes using them instead of simpler programs amounts using a sledgehammer to crack a nut.

Such design processes are key to creating real value with AI. They permit to shift the general question from "what is our place as humans in the new world shaped by AI?" to "what is the place of AI in the world we want to build, in our human world, among the other living beings?". In local contexts, this implies that concerned communities commit to a thorough collective reflection upon the problem they are confronted with and the manner they expect AI to help tackle it. Here lies the core of the ethical and political exploration societal communities must conduct to provide fruitful inputs and regulation for AI development and use. Such reflections touch upon many different key topics. We won't develop here those of them that are already quite well discussed (such as inequalities, environmental issues, questions with privacy protection, issue of economic model and intellectual property, …).[67] Rather, we would like to shed some light on the importance of a background exploration of what it means to be human. As we have seen, machines and humans are not interchangeable in many respects. In some cases (such as when it comes to decision-making), delegating to AI systems means radically removing humans doing to replace it by something far from equivalent. This raises the particularly acute challenge of discerning how to position AI for it to preserve or even serve the flourishing of human core specificities.

Exploring how to assist and support humans in their relationship to knowledge and truth (recommendation 8)

[67] See for instance: OECD, *Measuring the Environmental Impacts of Artificial Intelligence Compute and Applications: The AI Footprint*, OECD Digital Economy Papers no. 341, OECD Digital Economy Papers (2022), CCCXLI, doi:10.1787/7babf571-en; International Labour Organization and United Nations, *Mind the AI Divide: Shaping a Global Perspective on the Future of Work* (2024) <https://www.ilo.org/publications/major-publications/mind-ai-divide-shaping-global-perspective-future-work> [accessed 1 December 2025]; OECD, *Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions* (2025), doi:10.1787/795de142-en; UNESCO, *AI and Education: Protecting the Rights of Learners* (2025), doi:10.54675/ROQH4287; IEA, *Energy and AI – Analysis* (2025) <https://www.iea.org/reports/energy-and-ai> [accessed 1 December 2025].

As we developed, one of the core traits of humans is their fallible ability to relate to truth, using their intelligence to build corpuses of beliefs they judge deserving the title of knowledge. Putting AI at the genuine service of human intelligence is far from straightforward and will ask for deep discernment efforts to determine where and how AI can contribute positively to (or can on the contrary undermine) human efforts to elaborate knowledge and relate to truth.

A first key endeavor is to develop a robust individual and collective sense allowing to discern whether and how a given AI system in a specific context can be a source of knowledge. The clarifications and reminders given in recommendations 6 and 7 are crucial in this respect. There is no magical guaranty AI systems will produce valid results deserving to be held as pieces of knowledge. AI cannot be self-justificatory. Only humans have the autonomy and the ability to judge reasons available in favor of a (human or AI) production are good enough. In these matters, division of labor and delegation are common and indispensable. In many components of information technology, we (end users) trust subgroups of experts to assess the tools we are using and the reliability of their results. Although we can see by ourselves when our computer or the internet crashes, we largely delegate and trust IT and telecommunication companies to provide reliable devices. Similarly, we trust software companies to provide efficient word processors or spreadsheets, among many others. Depending on the content we access to thanks to digital technologies, we also usually trust content providers to share verified information (online journals, encyclopedia, …). We expect all these people and groups of people we trust to do their duty, to guarantee that systems they are involved with produce genuine knowledge.

It is interesting in this respect to note that generative AI constitutes an exception. Let's recall it again (better safe than sorry), the validity of LLMs' outcomes is not warranted. There is no specific subgroup of humans that is in charge (has the duty) of checking the singular content that is delivered to a given end user (only some samples are tested, especially during the training phase). This makes a decisive difference by comparison with what we normally expect and get, at least from providers and content we trust (for instance a reliable encyclopedia). In these cases, someone had an experience when elaborating and-or assessing a given content, a cognitive experience through which he or she tried to produce genuine knowledge, living a genuine relationship to truth. It is this type of experience we trust when delegating reliability assessment to other human beings. Seen in this way, what humans achieved in terms of knowledge elaboration and sharing, through their cooperation and with the support of information technology, is absolutely astonishing. Not just so many different contents, on so many different topics, accessible to almost anybody. But so many deep lived experiences of knowledge elaboration and validation that are put in common. This is the marvelous key point. And this is an element AI cannot reproduce. Instead of a strong shared and collaborative network of genuine cognitive work and experiences of relation to truth that enable legitimate trust building, LLMs provide end users with outcomes that are often true but that are never warranted (by a human). The burden of assessing these outcomes is shifted, whether they're aware of it or not, on end-users.

Of course, AI technology does not reduce to LLMs and generative AI. Many digital and AI tools are judged reliable and used in a large variety of contexts, even (or especially) in scientific ones.

But what we said permits to highlight the importance of AI literacy. End users must be aware of the strengths and limitations of the systems they are using and of the roles these systems can play or not in the patterns of knowledge building. End users have their part to play to preserve and prolong the efforts to relate to truth. To do so, they must be able to discern whether they can assume results produced are already pieces of knowledge (because they trust other humans who warrant them) or whether the assessment work remains to be done. It would be very dangerous to presuppose this ability is already enough developed. On the contrary, empirical studies suggest the opposite (laypersons tend to overestimate the reliability of LLM).[68] It is therefore very important to foster AI literacy in link with these epistemic issues. And the point here is not to say that systems we cannot trust as reliable sources of knowledge should be banned by principle. It can be very clever to mobilize generative AI to assist us in our exploratory tasks even when we know we cannot blind trust their results, as long as we refrain from relying integrally on them when it comes to justifying something deserves to be considered as a genuine piece of knowledge.[69] Our point rather that we must develop our capacity to attribute its right place to AI technology in our knowledge elaboration processes and in the ways we deal with our relationship to truth. This means being lucid about the current state of AI technologies, but also to explore the type of new systems we may develop to bring additional dedicated support in the various facets of our cognitive and epistemic lives.

Reflecting upon the manner AI can serve human (collective) intelligence (recommendation 9)

Fostering our ability to correctly assess the legitimacy of considering particular AI systems as sources of knowledge is a crucial dimension. However, it should not exhaust the discernment reflection. Beyond the question of the reliability of the tools, there is a more global question concerning the contribution of AI to the preservation and the development of human (individual and collective) intelligence. This is at the same time one of the overarching purposes AI should serve (the one we focus upon in this section) and a prerequisite for developing useful programs and aptly using them (as we just saw). In this more global perspective, the reliability of an AI systems in terms of knowledge production or transmission is no absolute warrant that it will contribute to the flourishing of humans' intellectual and cognitive life. In fact, a growing body of evidence suggests that overuse of generative AI systems such as LLMs can lead to deskilling or can impede cognitive development.[70] While it may well be acceptable or even desirable to quit doing some tasks, delegation to (generative) AI should always come with a thorough analysis of the skills we may lose or not develop and with discernment about the

[68] Steyvers and others, 'What Large Language Models Know and What People Think They Know'; Union (EBU), *News Integrity in AI assistants*.

[69] This is the idea behind the distinction, classical in epistemology, between the contexts of discovery and of justification. See for instance: Ladyman, *Understanding Philosophy of Science*, sec. 3.3.

[70] Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan, 'Is It Harmful or Helpful? Examining the Causes and Consequences of Generative AI Usage among University Students', *International Journal of Educational Technology in Higher Education*, 21.1 (2024), p. 10, doi:10.1186/s41239-024-00444-7; Nataliya Kosmyna and others, 'Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task', arXiv:2506.08872, version 1, preprint, arXiv, 10 June 2025, doi:10.48550/arXiv.2506.08872.

dispensability of these skills (not only to keep being able to verify AI outcomes, but also as some tasks and skills may be involved in other domains of people's cognitive development).

AI contribution to the preservation and development of human intelligence must also be discussed more globally, defocusing from generative AI and LLMs. As already evoked, digital and AI tools cannot be considered as mere instruments we could decide whether to integrate or not within our intellectual and cognitive processes. This is true enough in some cases but cannot be generalized. In fact, digital technology also became a milieu within which we live our cognitive lives and conduct or knowledge elaboration efforts. Based on powerful predictive algorithms, recommendation and search engines editorialize for us the vast amount of information available on the internet. More and more, generative AI is mobilized to organize and summarize contents that are too heavy to be processed directly. Thereby AI systems become a kind of 'cognitive extension' of human minds. They 'increasingly shape the informational substrate upon which human cognition operates'.[71] Here, the action of AI systems remains unnoticed, for a large part. Their outcomes in this domain will thus constitute a precondition, a more or less fertile ground for human intelligence.

As we have recalled in recommendation 7, the processes humans deploy to build knowledge, as well as the manner they relate to truth are far from infallible. Humans must strive to make responsible use of their freedom of thinking, to judge, from within their situated lived and embodied experience, whether a given belief comes with reasons that are good enough or not to be considered as knowledge. An important part of this lived relationship to truth and knowledge is collective. Knowing is not just convincing oneself, in isolation, that something is valid. It is also committing to the validity of what we hold as knowledge in front of others. When we think we know something, we expect others to agree with us. It matters whether they agree with us. If they don't it's a good reason to doubt. This is at the core of the elaboration of scientific knowledge (with scientific communities organized to favor such collective processes) but concerns also (first and maybe foremost) judgments and commitments with respect to basic and factual evidence pertaining to common sense (how could scientific communities properly function without such fundamental ground). A huge part of the quality of humans' intellectual and cognitive life rests upon that 'common decency', that basic will of humans to answer to the call of judging and knowing in common.[72]

In this regard, we must thus wonder whether the contributions of AI systems as a cognitive extension or a preconditioning milieu are positive or not. Are we provided with the most useful information and pieces of knowledge possible? Do AI systems foster mutual understanding and enrichment? Are they at the service of a genuine human collective effort to relate to truth and produce knowledge? Although the positive potential of AI is undeniable and particularly rich, some already effective contributions of AI systems to the structuration of our collective cognitive lives are extremely worrying. As is know well established, automatized editorialization of our informational landscapes often leads to cognitive bubbles, echo chambers where

[71] Massimo Chiriatti and others, 'System 0: Transforming Artificial Intelligence into a Cognitive Extension', *Cyberpsychology, Behavior, and Social Networking*, 28.7 (2025), pp. 534–42, doi:10.1089/cyber.2025.0201.

[72] Revault d'Allonnes, *La faiblesse du vrai*.

polarization grow wild and unweave the human epistemic net. This led some scholars to propose the concepts of epistemic harm and epistemic injustice consisting in the illegitimate degradation, because of digital technologies, of people's 'epistemic standing' (the perception by oneself and others of one's ability to know, to interpret and to faithfully testify).[73] In such post-truth contexts, trust and benevolence required to preserve and foster a fruitful collective cognitive life are discarded and undermined.

Such toxic contributions of AI systems are not a fatality. They are not intrinsic to AI employed to organize our cognitive milieu. AI possesses tremendous potential to foster collective intelligence. It could recommend us personalized information we need and that can enlarge our perspectives, making us more amenable to fruitful encounters with diverging opinions and ideas. It could help us spend more time in genuine relationships with others (notably helping us shifting away from our screens from time to time). However, these promising prospects clash with the reality of the "free" economic model that largely dominates the digital technology sector (at least in terms of software). In fact, this economic model relies on the capture of people's attention, an objective that presides to the design of AI systems preconditioning our cognitive milieu and is largely responsible for the acute problems we just mentioned.[74] Recent simulations suggest that it is not something that can be regulated from the outside, the problems being intrinsically connected with the basic objective of attention capture.[75] This means that to get the most of what AI can bring as a cognitive extension, we will need a lot of collective efforts and intelligence to re-orientate our economic models and consumer practices. Here again we see the deep political dimension of AI and the collective responsibility that comes with it.

Exploring how AI can contribute to human agency and responsibility (recommendation 10)

With the growth in complexity of digital technologies (and related socio-technical systems), intense discussion has emerged pointing to possible problems for attributing (moral or legal) responsibility in case of harm or problem generated by advanced systems (especially those based on machine learning techniques). Much of the debate relies on the shared acknowledgement that AI and digital technologies cannot be said to decide or act in any strong sense involving responsibility and accountability (as we recall in recommendations 6 and 7). Then, if AI cannot be responsible while it nevertheless complexifies and obfuscates the causal patterns, some responsibility gaps may occur, preventing responsibility attribution.[76]

---

[73] John Symons and Ramón Alvarado, 'Epistemic Injustice and Data Science Technologies', *Synthese*, 200.2 (2022), p. 87, doi:10.1007/s11229-022-03631-z; Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed, 'Epistemic Injustice in Generative AI', in *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (AAAI Press, 2025), pp. 684–97.

[74] Bronner, *Apocalypse cognitive*.

[75] Maik Larooij and Petter Törnberg, 'Can We Fix Social Media? Testing Prosocial Interventions Using Generative Social Simulation', arXiv:2508.03385, preprint, arXiv, 5 August 2025, doi:10.48550/arXiv.2508.03385.

[76] Andreas Matthias, 'The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata', *Ethics and Information Technology*, 6.3 (2004), pp. 175–83, doi:10.1007/s10676-004-3422-1.

Although these possible difficulties for attributing responsibility are highly significant (especially from a legal perspective), we would like here to approach them in an indirect way better align with the core of our discussion. In fact, some authors highlight that an important question in the background is the one of control over AI systems. 'The real problem lies elsewhere: autonomous machines should be built so as to exhibit a level of risk that is morally acceptable. If they fall short of this standard, they exhibit what we call 'a control gap.''[77] Beyond the possibility of responsibility attribution, it is of primary importance to ensure 'meaningful human control'.[78]

It is therefore of primary importance to foster AI literacy to cultivate the capacities of people and communities to correctly assess the type of tools they mobilize. What matters is to know which system can be trusted for what to properly discern between valid and fruitful use cases and dangerous ones. Automation or strong delegation should occur only with systems we can trust, either because end users tested them themselves or (most likely) when they trust subgroups with the assessment. Systems that cannot be trusted enough (as generative AI and LLMs) must always be used under direct human supervision or in contexts where invalid outcomes are not problematic. The same type of capabilities is necessary to contribute to deciding the tools and technology we should develop in the future. It is for instance crucial that enlarged communities participate in the reflection upon the specific problem of the loss of control over most advanced AI systems. To enable this, one must go beyond the usual prophecy about AI becoming more intelligent than humans. As we said, the problem is rather to ensure that complex and powerful mechanical systems do not become too unpredictable and misaligned with our objectives and values (it directly pertains to the 'meaningful human control' issue discussed here). And it is also about not deploying unpredictable systems in critical contexts. In this perspective, more discussion should bear upon the new trend of 'agentic AI' where generative AI (LLM or alike) are not restricted to text (or similar content) production anymore but can execute an enlarged range of actions (steering other programs to automatize agenda and appointment setting, purchase, messaging …), up to code compiling and execution. In such new configurations, consequences of malfunctioning and misalignment can become extremely dangerous.

More globally, ensuring meaningful human control will always be dependent upon the various concerned actors being able to exert their critical thinking and their capacity of decision-making in the strong sense (rooted in the possibility to sidestep, to be confronted with and arbitrate among a plurality of options …). Here we see emerging again an important pattern in the perspective of AI ethics and AI regulation: AI technology can strongly influence (positively or not) human abilities that are key for ensuring its own adequate development and use.

---

[77] Frank Hindriks and Herman Veluwenkamp, 'The Risks of Autonomous Machines: From Responsibility Gaps to Control Gaps', *Synthese*, 201.1 (2023), p. 21, doi:10.1007/s11229-022-04001-5.

[78] Filippo Santoni de Sio and Giulio Mecacci, 'Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them', *Philosophy & Technology*, 34.4 (2021), pp. 1057–84, doi:10.1007/s13347-021-00450-x.

One element is particularly worrisome in this respect. It seems ever clearer that AI technology comes with a temptation to offload the burden of acting and deciding.[79] For a part, this temptation can relate to a kind of promethean shame[80] that leads people to think AI systems are more capable than they are. A recent empirical study highlighted that people tend to rely more on (generative) AI when they do not feel confident in handling themselves.[81] In this context, one must wonder about the place we want to grant to risk taking and therefore to human mistakes. Obviously, we must strive to limit our errors and their consequences. But we should not let our discernment being integrally driven by a kind of blind and monolithic aversion for mistakes. The risk even exists to delegate some tasks to AI even when we perceive some limitations in the tool because we want to avoid taking responsibility for our doing. It may be tempting not to oppose an AI outcome even when it seems problematic because, in case there is a problem, 'it's the machine's fault'. This risk is particularly acute for high stakes decision-making, such as in the medical context. It is also interesting to recall that such an offloading is incompatible with the preservation of human creativity and margin of maneuver that is necessary to accompany the use of algorithmic systems automating decision-making through predictive analytics based on past data. Only such an autonomous and creative accompaniment can ensure that people are not subjected to algorithmic processing that denies their own creative nature, their own possibility to sidestep and change the manner they act and live.

Overall, offloading may respond to the temptation of reducing oneself to an inertial object,[82] a purely cybernetic being striving to minimize the efforts it deploys in search for libidinal satisfaction.[83] In fact, the capability of sidestepping, of imagining things could be different from what they actually are and in trying to influence the course of events is core to what it means to be human. But it is an extremely demanding capability, one we should cultivate and protect, possibly through a 'universal declaration of the rights of the human mind'.[84] More than a risk of a sudden loss of control over AI as with 'the abrupt takeover scenarios commonly discussed in AI safety', humanity may well be confronted with the danger of a 'gradual disempowerment', of an 'incremental erosion of human influence' that could lead to an 'irreversible loss of human influence over crucial societal systems, precipitating an existential catastrophe through the permanent disempowerment of humanity'.[85]

---

[79] Evan F. Risko and Sam J. Gilbert, 'Cognitive Offloading', *Trends in Cognitive Sciences*, 20.9 (2016), pp. 676–88, doi:10.1016/j.tics.2016.07.002; Michael Gerlich, 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking', *Societies*, 15.1 (2025), p. 6, doi:10.3390/soc15010006.

[80] Günthe Anders, Günther/Dries Anders Christia, and Christopher John Müller, *The Obsolescence of the Human* (University of Minnesota Press; Univ Of Minnesota Press, n.d.).

[81] Hao-Ping (Hank) Lee and others, 'The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers', *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA), CHI '25, 25 April 2025, pp. 1–22, doi:10.1145/3706598.3713778.

[82] Jean-Michel Besnier, *N'être plus qu'un objet: la tentation d'oublier la vie*, Technologia (Hermann éditeurs des sciences et des arts, 2025).

[83] Mark Hunyadi, 'La bataille de l'esprit', *Esprit*, no. 4 (April 2025), pp. 43–53, doi:10.3917/espri.2504.0043.

[84] Mark Hunyadi, *Déclaration universelle des droits de l'esprit humain: une proposition* (PUF, 2024).

[85] Jan Kulveit and others, 'Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development', arXiv:2501.16946, preprint, arXiv, 29 January 2025, doi:10.48550/arXiv.2501.16946.

Again, a lot of human efforts, intelligence and autonomy will be required to orientate the development and use of AI systems that will support and empower us in our decision-making activities, that will preserve and foster our autonomy and our decision-making abilities. How to encourage the development of algorithms that limit the temptation of illegitimate offloading? What type of decision-making do we want to delegate to machines, knowing that it will ultimately mean replacing decision-making with automation? What is the place of AI assistants and advisors in our processes of (moral) decision-making? What is the place we grant for human mistake? What is the price we are ready to pay to defend and even cultivate human autonomy? These are very acute questions that should be discussed in as large as possible communities of concerned persons.

Problematizing the notions of progress, good life and vulnerability (recommendation 11)

In the background of the discussion developed up to now lies the key question of the good life. What does it mean to live a good life? What does it mean to improve our lives? What is the connection with the ideas of progress and innovation, especially with AI technology? In these matters, one can hardly uncritically adopt radical techno-optimists or techno-solutionists position of the type defended by some powerful actors of the Silicon Valley such as Peter Thiel or Marc Adreessen according to which: 'there is no material problem – whether created by nature or by technology – that cannot be solved with more technology' or that 'we are poised for an intelligence takeoff that will expand our capabilities to unimagined heights', with AI considered as 'our alchemy, our Philosopher's Stone – we are literally making sand think'.[86] These sorts of narratives assume a principled causal link between innovation and progress in technology and genuine improvement of our lives.

Adequately steering the development and use of AI necessitates to resist this type of techno-solutionists shortcuts. It is particularly important to rather always be ready to refine and enlarge our understanding of what "genuine human" progress means. Before looking for technological solutions to a given problem, we must ensure our analysis of it is deep and rich enough. For instance, it is far from obvious that we can properly answer to the extreme exhaustion of healthcare professionals by just providing them with tools enhancing their efficiency and productivity or by replacing them in some tasks by robots or other automata. These solutions can only be legitimate as outcomes of a broad collective reflection (making room to the realities of the situated practices) on the root causes of overburdening and lack of personnel. Similarly, it cannot suffice to answer to the growing feeling of loneliness that touches an ever-increasing number of people by artificial companions and social robotics. In such cases, the technological solution can even become a manner of perpetuating and even worsening a deeper problem.

[86] Andreessen, 'The Techno-Optimist Manifesto'; See also Pieter Thiel's interview for The New York Times: Ross Douthat, 'Opinion | Peter Thiel and the Antichrist', Opinion, *The New York Times*, 26 June 2025 <https://www.nytimes.com/2025/06/26/opinion/peter-thiel-antichrist-ross-douthat.html> [accessed 24 October 2025].

Such caution and refinement in the analysis of the connection between technological progress and innovation on the one hand and genuine human progress on the other is at the core of the technocritic tradition we evoked in the introduction. One could easily relate to this school of thought the words of Pope Francis alerting against our tendency to make a (morally) blind use of our (technological) powers assuming that more power amounts necessarily to human progress (often reducing our understanding of the notion to matters of utility or security presented as imperatives).[87] This tendency culminates in what Francis calls the 'technocratic paradigm' that 'exalts the concept of a subject who, using logical and rational procedures, progressively approaches and gains control over an external object' in a spirit of 'possession, mastery and transformation'. With the technocratic paradigms humans tend to consider reality they intervene in as 'something formless, completely open to manipulation' from which to extract as much as possible (instead of 'being in tune with and respecting the possibilities offered by the things themselves').[88] Within such a thinking environment marked by the domination of instrumental rationality, it becomes difficult to conceive of progress other than in terms of efficiency and measurable performance. Problems humans strive to address, and even human affairs in general tend to be reduced to indicators to optimize. Human themselves become mere resources, skills and roles or functions, 'which can then be duplicated, improved, surpassed'.[89]

Not only can this type of mindset lead to promote or caution mechanistic and algorithmic forms of governance,[90] it also nurtures the risk of mutilating the legitimate human search for freedom and emancipation, reducing it to an obsessional rejection of all limits. In this perspective, any weakness, any vulnerability, any possibility of failure is a defect one must strive to correct. Such a view of human development as progression toward perfection and unlimited might is highly problematic for multiple reasons, not the least of which being its elitist and inequalitarian dimension.[91] Here, we would like to focus upon epistemological, moral and anthropological reasons to take distance with this mutilated conception of human progress.

As we have seen in the previous sections (recommendations 8 to 10), the very possibility of mistake and failure is intrinsic to knowledge and decision-making. Trying to reduce the number of mistakes we make is a duty and constitutive of knowledge and moral decision-making. But arguing that we should fight against the very possibility of failure and mistake is a totally different thing that does not amount to the improvement of knowledge elaboration and decision-making, but to their eradication. As we have seen, knowledge and decision-making in the strong sense imply a genuine margin of maneuver irreducibly including the possibility of choosing wrong. Improving our knowledge and our decisions does not necessitate reducing

---

[87] Pope Francis, 'Laudato Si' (24 May 2015)', 24 May 2015, para. 105 <https://www.vatican.va/content/francesco/en/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html> [accessed 5 December 2025].

[88] Pope Francis, 'Laudato Si' (24 May 2015)', para. 106.

[89] Our translation, Nicolas Léger and Adrien Tallent, 'L'IA aux frontières de l'esprit. Introduction', *Esprit*, no. 4 (April 2025), pp. 35–42 (p. 36), doi:10.3917/espri.2504.0035.

[90] Antoinette Rouvroy, 'Algorithmic Governmentality', in *More Posthuman Glossary*, ed. by Rosi Braidotti, Emily Jones, and Goda Klumbytė (Bloomsbury Academic, 2022), doi:10.5040/9781350231467.

[91] Michael J. Sandel, *The Case against Perfection: Ethics in the Age of Genetic Engineering* (Harvard University Press, 2009), doi:10.2307/j.ctvjz80mc.

this margin of maneuver in an illusory quest for infallibility. Rather it demands refining our collective sense of responsibility, our ability to exert our critical thinking and common decency. Improving in this domain is a life path, a commitment to do our best (individually and collectively) to decide as best as we can without absolute warrant, to act and believe for good (though fallible) reasons.

The same can be said of vulnerability in a broader and even more fundamental sense. Absolute robustness of the body and the mind does not constitute the perfection of human life but its radical negation. Again, it is our duty and responsibility to do our best to cure or prevent injuries and diseases. It is also more than legitimate to try to avoid hurting people and being hurt by them. But this is not the same as trying to eradicate as much as possible the possibility of being hurt, of getting sick or even of dying. We can suffer and be injured because we are vulnerable. But it would be a dangerous mistake to reduce vulnerability to these negative aspects only. As David Doat puts it, '[v]ulnerability is not weakness or poverty. Nor can it be reduced to old age, disability or illness. (…) we need to distinguish between "vulnerability" and "vulneration". The former refers to the possibility of being affected in one's physical or psychological structure; the latter refers to the state following an injury. It's important to make the difference. During a romantic encounter, for example, the lovers are in a state of vulnerability as they expose themselves to each other, each allowing themselves to be affected by the beloved, but both are not injured.'[92]

To summarize, the notion of genuine human progress cannot reduce to an increase in power that would lead closer to perfection, infallibility and invulnerability. Everything that counts in our lives comes at the price of vulnerability and fallibility. It is because we have the ability to choose that we can make mistakes. It is because we can judge the quality of reasons to believe something and imagine alternatives that we can elaborate knowledge, but this irreducibly implies the possibility of erroneous assessments. It is because we are alive that we can get negatively affected, injured and traumatized. It is because we can feel joy and love that we can also feel sadness and despair. Wondering whether and how AI can make us all powerful, infallible and invulnerable is thus the wrong question. Rather we should reflect upon how to develop and use AI systems that could help us better tame and balance the ambivalent but essential vulnerability that lies at the deepest heart of who we are.

Cultivating our sensitivity to life and conscious lived experience (recommendation 12)

What we just said leads us to a last topic we would like to explore in the perspective of reinforcing our ability to correctly orientate the development and use of AI. It concerns the legitimacy and importance of valuing life and lived experience at the core of which lie affectability and vulnerability. These dimensions are essential to genuine relationships between humans and more broadly between living beings. In this regard, we must warn again against

[92] Extract from Brigitte Bègue, 'La vulnérabilité peut être une chance. Mais on l'oublie', interview with David Doat, 5 March 2021, Actualités sociales hebdomadaires N° 3199 <https://www.ash.tm.fr/hebdo/3199/entretien/la-vulnerabilite-peut-etre-une-chance-mais-on-loublie-634607.php> [accessed 5 December 2025], our translation.

the "outcomist" tendency that tends to discard these dimensions as not objectifiable or demonstrable, urging us to focus on outcomes only. Although we spent several sections discussing and defusing this stream of reasoning (see recommendation 7), its possible toxic consequences on our ability to build AI technology at the genuine service of humanity.

Based on outcomism, some argue in favor of the possibility of AI consciousness (in a foreseeable future) or more perniciously claim that trying to differentiate humans and machines on this aspect is meaningless. There is no point wondering about AI being genuinely conscious beyond its observable behaviors and outcomes. If we have the feeling it is conscious because it behaves convincingly, it is conscious … there nothing more to say. The important thing is the manner it makes us feel.

According to such a line of thought, many uses of AI for social relationships become more easily admissible. Some empirical evidence suggests that people, when presented in blind settings with written medical communications produced by LLMs or by actual healthcare professionals, tend to prefer those produced by the machine, especially for its more empathetic content.[93] If people prefer these answers, why not giving them what they prefer. Similarly, if people like AI-generated pieces of art (at least when we do not insist too much on their origin), why should we refrain from producing music, pictures or other content this way. If people consider AI companions as true friends that can bring them social relationships and actual recomfort, why depriving them of this effective wellbeing.

As we have seen, there is no compelling reason to admit outcomism. On the contrary, there are many ways to argue in favor of our basic and traditional intuitions about what type of entities are alive or not, are capable of consciousness and affectability in the psychological sense, or can enter in genuine types of relationships. This means we are perfectly legitimate in valuing not only the quality of outcomes we are presented with, but also the manner they have been produced, and especially the presence of lived experience, vulnerability and affectability, the presence of a genuine person upstream. To properly orientate the development and uses of AI systems, we must therefore cultivate our sensitivity to life and to genuine vulnerable and affectable persons. More precisely, we must foster and cultivate our ability to value their presence (and not just the appearance of such presence), to assess when their presence has decisive value.

A first domain in which the presence of genuine lived experience is indispensable is the one of relationship to knowledge and truth as well as the one of decision-making. No knowledge elaboration or decision-making in the strong sense without one or several genuine persons making more or less responsible use of the freedom they have from within their lived experience. In many other fields, knowing there is someone in front of us (rather than having mere appearances) is key. It is especially the case in healthcare where empathy and genuine doctor-patient relationships play a crucial role for therapy and recovery.[94] The same could be

[93] The experiment being made with answers to questions on healthcare forums. See: John W. Ayers and others, 'Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum', *JAMA Internal Medicine*, 183.6 (2023), pp. 589–96, doi:10.1001/jamainternmed.2023.1838.

[94] Joan C. Tronto, *Caring Democracy: Markets, Equality, and Justice* (New York University Press, 2013).

said of the importance of human relationships in psychotherapeutic processes. The presence of a genuine person we talk to during therapeutic conversations is indispensable,[95] even if it may feel more demanding than merely talking to a chatbot.

In fact, genuine human relationships put us at risks, come with their share of discomfort, insecurity and worry. There lies a specific risk with social AI and artificial companions. In most cases, AI systems of this type are designed as products at our service. They will never strongly oppose our will. By contrast with the anguish genuine relationships may trigger, it may feel very attractive to get a very convincing appearance of relationship but purged from any risk of being rejected, abandoned, judged or hurt. As tempting as this may be, it would nevertheless amount to removing everything that makes a genuine human relationship, removing any value to this imitation of acceptance and love. How can it truly be love or acceptance if it is not on the background of a genuine lived experience of a free person that could not give (could have not given) it?[96] Empirical evidence seems to confirm such limitations of AI companions when used in a massive way to compensate for social isolation.[97]

The value of the presence of a genuine person can be made plainly visible when considering the expression of words of compassion, for instance in front of a dying person. The words themselves (or other means of expressions) are not the most important in such a situation. What is primary is that they signal a genuinely experienced feeling of compassion. The same type of thing could be said in the case of a child making us a drawing. In this context, the 'objective' esthetic quality of the outcome (the drawing) in itself is largely secondary. What would be the value of an objectively very nice picture the kid gives us, but which would be produced in few minutes through a generative AI program? Could it be compared with the value of an imperfect drawing (we are not even sure what it may represent) the kid spent several hours realizing, putting a lot of effort and intention into it? We don't really care here about the 'objective' esthetic quality of the picture. What matters is the thick lived experience it results from.

In a way, things can be a bit in between in the case of 'professional' art. The 'objective' aspect of the piece of art may legitimately matter. In most cases, we do not take into account only or primarily the lived experience of the artist when he or she elaborated the artwork. But we do also take such dimensions into account. We are legitimate to integrate in our esthetic judgement our knowledge about where the piece of art comes from, about who made it. More than this, it is perfectly legitimate that knowing it has been made by a genuine person instead of an AI system makes us feel the artwork differently. Undermining or denying this legitimacy (especially based on flawed epistemological arguments about the obligation to be objective) constitutes a grave aggression against core dimensions of what it means to be alive and to be

[95] Jana Sedlakova and Manuel Trachsel, 'Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?', *The American Journal of Bioethics*, 23.5 (2023), pp. 4–13, doi:10.1080/15265161.2022.2048739.

[96] AI Research Group of the Centre for Digital Culture, 'Encountering Artificial Intelligence', p. 117.

[97] Cathy Mengying Fang and others, *How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Controlled Study* (MIT Media Lab and Open AI, 2025) <https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/> [accessed 29 November 2025].

human. We must learn to oppose strong resistance to such aggression by (let's say it again) fostering and cultivating our sensitivity to life and to genuine human presence.

Rooted in such a robust background, we can wonder about the desirable types of AI systems and uses, namely those that will contribute to fostering our sensitivity to life and to cultivating our ability to enter in genuine relationships. In general, it is indispensable to always disclose when an outcome is produced by an AI system instead of a human. As we just saw, knowledge of this is necessary for people to evaluate what they are presented with. Beyond this, we must deploy serious efforts to identify what could be truly positive uses of systems that can convincingly mimic humans or other living beings, as well as to determine what features such systems should possess or not. This is a very difficult question.

Take for instance the case of AI companions. We may consider that we could use them as sophisticated toys or support for entertainment, as long as we know they are not human. After all, AI companions may be compared to a good movie, a good book, or, even better a good video games with characters in them that we like to follow in their adventures and possibly interact with. There could be here truly positive or innocuous uses. One must nonetheless pay attention to a delicate point: it may not always be enough to just being clearly aware that we interact with a human mimicking AI system and not with a genuine person. In fact, AI companions push at its extreme the attribution of human appearance to artifacts designed to serve users and consumers. AI systems reaching convincing levels in such matter raises the risk of 'both schooling its users in the negation of the other and fostering a culture that absorbs intimacy into a schema of property relations and rights rather than into the vulnerable gift of true intersubjectivity'. 'Where there is no "other," but only the *appearance* of an other at our disposal, concurrent with the absence of the demand that would be exercised upon one's own self-gift by confrontation with a true other, we risk being conditioned in a dangerous talent for exploitation.'[98]

We thereby risk fostering the progressive reduction of human relationships to service interactions, with the danger of becoming increasingly less able to tolerate the true autonomy of the others, the autonomy that makes relationships genuine ones. Instead of being capable of seeing frictions and opposition as also opportunities for genuine encounter with true persons, we may start considering people responsible for these resistances as faulty humans (as we would do with malfunctioning artifacts). Such issues may lead to open a debate parallel to the discussions about AI welfare and rights, but for different reasons. In fact, it might become necessary to regulate the manner we interact with artificial companions or assistants, not because we could hurt them through inappropriate behaviors but because we could harm ourselves.[99]

We can thus see that a lot of reflection, exploration and effort will be required to make room for life and lived experience in the contexts where their value is primary. While it may be easy and tempting to use AI systems to propose (pale) substitutes to social interactions, putting AI at the service of the intensification of genuine relationships and their quality constitutes a

[98] AI Research Group of the Centre for Digital Culture, 'Encountering Artificial Intelligence', p. 120.
[99] AI Research Group of the Centre for Digital Culture, 'Encountering Artificial Intelligence', pp. 128–30.

demanding challenge. Digital technologies (especially social networks and now social AI) may give a superficial impression of social proximity while it in reality contributes to isolation and alienation.[100] The same sort of risks may occur depending on the manner we will react to empirical findings evidencing the impression of empathy generative AI can trigger. They sometimes appear more empathetic than humans, as we mentioned with answers on healthcare public forums. Of course, the authors do not conclude that healthcare professionals should be replaced by AI systems. Rather, they suggest that LLMs could draft more empathetic communication material for these professionals (with the possibility to improve patients therapeutic trajectories and to reduce professionals overburdening).[101] More globally, it is often proposed, especially in this domain of healthcare, to use AI systems and advanced robotics to offload overburden professionals.

To our mind, the exploration can be worthwhile but should not be conducted without deep analysis of the causes behind overburdening or other problems such as difficulties in communication from (healthcare) professionals (in purely medical terms and with respect to empathy). We must seek the causes of such problems and identify the various options to mitigate them. It is far from obvious that AI support is the best option. Especially, outsourcing the production of apparently empathetic communication may even reinforce the exhaustion of healthcare professionals. Maybe they would like to take more time for genuine relationships with their patients but cannot because of the overburdening. Maybe their training should also be interrogated, notably with respect to a possible tendency to reduce living beings and human persons to biological functions to monitor, maintain or restore. Again, it is of primary importance to deeply reflect on the genuine purposes AI systems should serve, striving to keep at the center of discernment efforts key dimensions such as the value of life and conscious lived experience we discussed in this section.

Concluding remarks

Let's close this exposition with some general remarks and highlights on most important messages. First and foremost, it is important we insist upon the spirit of the collective endeavor for discernment we propose here. In no case can it be as simple as merely claiming that the human is always wonderful and should be preferred over the machine in every context. Of course, humans are fallible, they make mistakes, they can be particularly unpleasant to each other. Some humans can behave in a totally barbarian and inhuman way. Our central point is rather to say that the best way to move forward is not always to strive to make AI systems ensure functions where humans can be faulty, especially when it involves illusion of infallibility or mere appearance of kindness and benevolence. The best way to move forward cannot be to put humans and machines in competition. As we clarified, it is most of the time illegitimate to claim that AI can replace humans in some tasks. What is true is that AI can ensure certain functions that would demand a human presence otherwise. But this rarely means that the

[100] Sherry Turkle, *Alone Together: Why We Expect More from Technology and Less from Each Other*, Third edition, revised trade paperback edition (Basic Books, 2017).

[101] Ayers and others, 'Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum'.

replacement of humans by AI leads to an equivalent situation. We must always keep in mind that we replace a living person with all the richness of her lived experience by pure mechanism and automatism.

Put in a nutshell, the primary question is not to determine what AI can do better than humans (even if it is a legitimate question when approached in all its complexity without reduction on the manner we understand situations). Rather, the main question we should start from is: "how can AI technology support us in becoming better humans?". This is a key shift to adopt if we want to deploy robust AI ethics and regulation. The question with AI should not be: "what is our place as humans in the new world of AI?", as if AI technology was not the results of the choices and doing of (some) humans and the only thing human communities could do was to adapt to this new world. Rather, we should collectively wonder: "what is the place of AI technology in our human world, and with the other living beings?" "What is the contribution AI can bring to the development of more humane societies?". Framed this way, it becomes clearly visible that the question of AI technology development and use is (at least) as much a political and ethical issue as it is a technical question.

In fact, we cannot answer such questions without collectively exploring what it means to become better humans, what are the society projects AI should serve. This necessitates a considerable political commitment from the behalf of large portions of human communities. We need to discern together, to make us of our critical thinking and our ability to genuinely decide. In this respect, it is important to recall the circular threat AI raises. We need a lot of free attention time of good quality to adequately participate in these discernment efforts while recommendation algorithms are currently very efficient to siphon off this free attention time. We need to foster our ability to decision-making in the strong sense while generative AI opens tempting possibilities for excessive cognitive offloading. We must cultivate our sensitivity to life and to the importance of genuine lived phenomenal experience in an era of ever better mimicking machines. Without enough discernment, developed AI systems and uncritically adopted uses can undermine the capabilities we need to properly discern and build desirable AI technology. To nurture these discernment efforts, we need to foster better understanding as well of AI technology themselves as of human nature and condition. AI literacy is absolutely key to correctly grasp what we can reasonably expect or not from AI systems (especially in terms of reliability). In addition, we must always dig in the context of a difficulty we encounter and we consider mitigating with AI. We always have to wonder whether AI will bring a genuine solution to the problem or whether it can at best help us to temporally cope with a deeper problem we should address differently. Not doing seriously so would amount to take the risk of using AI to perpetuate acute human difficulties, to depriving us from the opportunity to better develop and flourish.

Again, it is legitimate to be enthusiast and optimistic with AI. We must be so. AI comes with tremendous potential to support us in our development and flourishing. But we must at the same time always foster and keep clear awareness of the price we are called to pay in terms of commitment to challenging efforts of ethical and political discernment.