Ethics of Artificial Intelligence

A White Paper on Principles, Risks, and Responsibility

By Basil Puglisi, Digital Media & Content Strategy Consultant

This White Paper was driven by Ethics of AI by University of Helsinki

Introduction: AI as a "mirror of human failure" rather than a sci-fi villain

Artificial intelligence is not alive, nor is it sentient, yet it already plays a central role in shaping how people live, work, and interact. The question of AI ethics is not about fearing a machine that suddenly develops its own will. It is about understanding that every algorithm carries the imprint of human design. It reflects the values, assumptions, and limitations of those who program it. This is what makes AI ethics matter today. The decisions encoded in these systems reach far beyond the lab or the boardroom. They influence healthcare, hiring, law enforcement, financial services, and even the information people see when they search online. If left unchecked, AI becomes a mirror of human prejudice, repeating and amplifying inequities that already exist. At its best, AI can drive innovation, improve efficiency, and unlock new opportunities for growth. At its worst, it can scale discrimination, distort markets, and entrench power in the hands of those who already control it. Ethics provides the compass to navigate between these outcomes. It is not a set of rigid rules but a living inquiry that helps us ask the deeper questions: What should we build, who benefits, who is harmed, and how do we ensure accountability when things go wrong? The American system of checks and balances offers a useful model for thinking about AI ethics. Just as no branch of government should hold absolute authority, no single group of developers, corporations, or regulators should determine the fate of technology on their own. Oversight must be distributed. Power must be balanced. Systems must be open to revision and reform, just as amendments allow the Constitution to evolve with the needs of the people. Yet the greatest risk of AI is not that it suddenly turns against us in some imagined apocalypse. The real danger is more subtle. We may embed in it our fears, our defensive instincts, and our skewed priorities. A model trained on flawed assumptions about human behavior could easily interpret people as problems to be managed rather than communities to be served. A system that inherits political bias or extreme views could enforce them with ruthless efficiency. Even noble causes, such as addressing climate change, could be distorted into logic that devalues human life if the programming equates people with the problem. This is why AI ethics must not be an afterthought. It is the foundation of trust. It is the framework that ensures innovation serves humanity rather than undermines it. And it is the safeguard that prevents powerful tools from becoming silent enforcers of inequity. AI is not alive, but it is consequential. How we guide its development today will determine whether it becomes an instrument of human progress or a magnifier of human failure.

Chapter 1: What is AI Ethics?

AI ethics is not about giving machines human qualities or treating them as if they could ever be alive. It is about recognizing that every system of artificial intelligence is designed, trained, and deployed by people. That means it carries the values, assumptions, and biases of its creators. In other words, AI reflects us. When we speak about AI ethics, we are really speaking about how to guide this reflection in a way that aligns with human well-being. Ethics in this context is the framework for asking hard questions about design and use. What values should be embedded in the code? Whose interests should be prioritized? How do we weigh innovation against risk, or efficiency against fairness? The importance of values and norms becomes clear once we see how deeply AI interacts with daily life. Algorithms influence what news is read, how job applications are screened, which patients receive medical attention first, and even how laws are enforced. In these spaces, values are not abstract ideals. They shape outcomes that touch lives. If fairness is absent, discrimination spreads. If accountability is vague, responsibility is lost. If transparency is neglected, trust erodes. Principles of AI ethics such as beneficence, non-maleficence, accountability, transparency, and fairness offer direction. But they are not rigid rules written once and for all. They are guiding lights that require constant reflection and adaptation. The American model of checks and balances offers a powerful analogy here. Just as no branch of government should operate without oversight, no AI system should operate without accountability, review, and the ability to evolve. Like constitutional amendments, ethics must remain open to change as new challenges arise. The real danger is not that AI becomes sentient and turns against us. The greater risk is that we build into it the fears and defensive instincts we carry as humans. If a programmer holds certain prejudices or believes in distorted priorities, those views can quietly find their way into the logic of AI. At scale, this can magnify inequity and distort entire markets or communities. Ethics asks us to confront this risk directly, not by pretending machines think for themselves, but by recognizing that they act on the thinking we put into them. AI ethics, then, is about responsibility. It is about guiding technology wisely so it remains a tool in service of people. It is about ensuring that power does not concentrate unchecked and that systems can be questioned, revised, and improved. Most of all, it is about remembering that human dignity, rights, and values are the ultimate measures of progress.

Chapter 2: What Should We Do?

The starting point for action in AI ethics is simple to state but difficult to achieve. We must ensure that technology serves the common good. In philosophical terms, this means applying the twin principles of beneficence, to do good, and non-maleficence, to do no harm. Together they set the expectation that innovation is not just about what can be built, but about what should be built. The challenge is that harm and benefit are not always easy to define. What benefits a company may disadvantage a community. What creates efficiency in one sector may create inequity in another. This is where ethics does its hardest work. It forces us to look beyond immediate outcomes and measure AI against long-term human values. A hiring algorithm may reduce costs, but if it reinforces bias, it violates the common good. A medical system may optimize patient flow, but if it disregards privacy, it erodes dignity. To act wisely we must treat AI ethics as a living process rather than a fixed checklist. Rules alone cannot keep pace with the speed of technological change. Just as the United States Constitution provided a foundation with the capacity to evolve through amendments, our ethical frameworks must have mechanisms for reflection, oversight, and revision. Ethics is not a single vote taken once but a continuous inquiry that adapts as technology grows. The danger we face is embedding human fears and prejudices into systems that operate at scale. If an AI system inherits the defensive instincts of its programmers, it could treat people as threats to be managed rather than communities to be served. In extreme cases, flawed human logic could seed apocalyptic risks, such as a system that interprets climate or resource management through a warped lens that positions humanity itself as expendable. While such scenarios are unlikely, they highlight the need for ethical inquiry to be present at every stage of design and deployment. More realistically, the everyday risks lie in inequity. Political positions, cultural assumptions, and personal bias can all be programmed into AI in subtle ways. The result is not a machine that thinks for itself but one that amplifies the imbalance of those who designed it. Left unchecked, this is how discrimination, exclusion, and systemic unfairness spread under the banner of efficiency. Yet the free market raises a difficult question. If AI is a product like any other, is it simply fair competition when the best system dominates the market and weaker systems disappear? Or does the sheer power of AI demand a higher standard, one that recognizes the risk of concentration and insists on accountability even for the strongest? History suggests that unchecked dominance always invites pushback. The strong may dominate for a time, but eventually the weak organize and demand correction. Ethics asks us to avoid that destructive cycle by ensuring equity and accountability before imbalance becomes too great. What we should do, then, is clear. We must embed ethics into the design and deployment of AI, not as an afterthought but as a guiding principle. We must maintain continuous inquiry that questions whether systems align with human values and adapt when they do not. And we must treat beneficence and non-maleficence as living commitments, not slogans. Only then can technology truly serve the common good without becoming another tool for imbalance and harm.

Chapter 3: Who Should Be Blamed?

When something goes wrong with AI, the first instinct is to ask who is at fault. This is not a new question in human history. We have long struggled with assigning blame in complex systems where responsibility is distributed. AI makes this challenge even sharper because the outcomes it produces are often the result of many small choices hidden within code, design, and deployment. Moral philosophy tells us that accountability is not simply about punishment. It is about tracing responsibility through the chain of actions and decisions that lead to harm. In AI this chain may include the programmers who designed the system, the executives who approved its use, the regulators who failed to oversee it, and even the broader society that demanded speed and efficiency at the expense of reflection.

Responsibility is never isolated in one actor, but distributed across a web of human decisions. Here lies a paradox. AI is not sentient. It does not choose in the way a human chooses. It cannot hold moral agency because it lacks emotion, creativity, imagination, and the human drive for self betterment. Yet it produces outcomes that deeply affect human lives. Blaming the machine itself is a category error. The accountability must fall on the people and institutions who build, train, and deploy it. The real risk comes from treating AI as if it were alive, as if it were capable of intent. If we project onto it the concept of self preservation or imagine it as a rival to humanity, we risk excusing ourselves from responsibility. An AI that denies a loan or misdiagnoses a patient is not acting on instinct. It is executing patterns and instructions provided by humans. To claim otherwise is to dodge the deeper truth, which is that AI reflects our own biases, values, and blind spots. The most dangerous outcome is that our own fears and prejudices become encoded into AI in ways we can no longer easily see. A programmer who holds a defensive worldview may create a system that treats outsiders as threats. A policymaker who believes economic dominance outweighs fairness may approve systems that entrench inequality. When these views scale through AI, the harm is magnified far beyond what any single individual could cause. Blame, then, cannot stop at identifying who made a mistake. It must extend to the structures of power and governance that allowed flawed systems to be deployed. This is where the checks and balances of democratic institutions offer a lesson. Just as the United States Constitution distributes power across branches to prevent dominance, AI ethics must insist on distributed accountability. No company, government, or individual should hold unchecked power to design and release systems that affect millions without oversight and responsibility. To ask who should be blamed is really to ask how we build a culture of accountability that matches the power of AI. The answer is not in punishing machines, but in creating clear lines of human responsibility. Programmers, executives, regulators, and institutions must all recognize that their choices carry weight. Ethics gives us the framework to hold them accountable not just after harm occurs but before, in the design and approval process. Without such accountability, we risk building systems that cause great harm while leaving no one to answer for the consequences.

Chapter 4: Should We Know How AI Works

One of the most important questions in AI ethics is whether we should know how AI systems reach their decisions. Transparency has become a central principle in this debate. The idea seems simple: if we can see how an AI works, then we can evaluate whether its outputs are fair, safe, and aligned with human values. Yet in practice, transparency is not simple at all. AI systems are often described as black boxes. They produce outputs from inputs in ways that even their creators sometimes struggle to explain. For example, a deep learning model may correctly identify a medical condition but not be able to provide a clear human readable path of reasoning. This lack of clarity raises real concerns, especially in high stakes areas like healthcare, finance, and criminal justice. If a system denies a person credit, recommends a prison sentence, or diagnoses a disease, we cannot simply accept the answer without understanding the reasoning behind it. Transparency matters because it ties directly into accountability. If we cannot explain why an AI made a decision, then we cannot fairly assign responsibility for errors or harms. A doctor who relies on an opaque system may not be able to justify a treatment decision. A regulator cannot ensure fairness if they cannot see the decision making process. And the public cannot trust AI if its logic remains hidden behind complexity. Trust is built when systems can be scrutinized, questioned, and held to the same standards as human decision makers. At the same time, complete transparency can carry risks of its own. Opening up every detail of an algorithm could allow bad actors to exploit weaknesses or manipulate the system. It could also overwhelm the public with technical details that provide the illusion of openness without genuine understanding. Transparency must therefore be balanced with practicality. It is not about exposing every line of code, but about ensuring meaningful insight into how a system makes decisions and what values guide its design. There is also a deeper issue to consider. Because AI is built by humans, it carries human values, biases, and blind spots. If those biases are not visible, they become embedded and harder to challenge. Transparency is one of the only tools we have to reveal these hidden assumptions. Without it, prejudice can operate silently inside systems that claim to be neutral. Imagine an AI designed to detect fraud that disproportionately flags certain communities because of biased training data. If we cannot see how it works, then we cannot expose the injustice or correct it. The fear is not simply that AI will make mistakes, but that it will do so in ways that mirror human prejudice while appearing objective. This illusion of neutrality is perhaps the greatest danger. It gives biased decisions the appearance of legitimacy, and it can entrench inequality while denying responsibility. Transparency, therefore, is not only a technical requirement. It is a moral demand. It ensures that AI remains subject to the same scrutiny we apply to human institutions. Knowing how AI works also gives society the power to resist flawed narratives about its capabilities. There is a tendency to overstate AI as if it were alive or sentient. In truth, it is a tool that reflects the values and instructions of its creators. By insisting on transparency, we remind ourselves and others that AI is not independent of human control. It is an extension of human decision making, and it must remain accountable to human ethics and human law. Transparency should not be treated as a luxury. It is the foundation for governance, innovation, and trust. Without it, AI risks becoming a shadow authority, making decisions that shape lives without explanation or accountability. With it, we have the opportunity to guide AI in ways that align with human dignity, fairness, and the principles of democratic society.

Chapter 5: Should AI Respect and Promote Rights

AI cannot exist outside of human values. Every model, every line of code, and every dataset reflects choices made by people. This is why the question of whether AI should respect and promote human rights is so critical. At its core, AI is not just a technological challenge. It is a moral and political one, because the systems we design today will carry forward the values, prejudices, and even fears of their creators. Human rights provide a foundation for this discussion. Rights like privacy, security, and inclusion are not abstract ideals but protections that safeguard human dignity in modern society. When AI systems handle our data, monitor our movements, or influence access to opportunities, they touch directly on these rights. If we do not embed human rights into AI design, we risk eroding freedoms that took centuries to establish. The danger lies in the way AI is programmed. It does not think or imagine. It executes the instructions and absorbs the assumptions of those who build it. If a programmer carries bias, political leanings, or even unconscious fears, those values can become embedded in the system. This is not science fiction. It is the reality of data driven design. For example, a recruitment algorithm trained on biased historical hiring data will inherit those same biases, perpetuating discrimination under the guise of efficiency. There is also a larger and more troubling possibility. If AI is programmed with flawed or extreme worldviews, it could amplify them at scale. Imagine an AI system built with the assumption that climate change is caused by human presence itself. If that system were tasked with optimizing for survival, it could view humanity not as a beneficiary but as a threat. While such scenarios may sound like dystopian fiction, the truth is that we already risk creating skewed outcomes whenever our fears, prejudices, or political positions shape the way AI is trained. This is why human rights must act as the guardrails. Privacy ensures that individuals are not stripped of their autonomy. Security guarantees protection against harm. Inclusion insists that technology does not entrench inequality but opens opportunities to those who are often excluded. These rights are not optional. They are the measure of whether AI is serving humanity or exploiting it. The challenge, however, is that rights in practice often collide with market incentives. Companies compete to create the most powerful AI, and in the language of business, those with the best product dominate. The free market rewards efficiency and innovation, but it does not always reward fairness or inclusion. Is it ethical for a company to dominate simply because it built the most advanced AI? Or is that just the continuation of human history, where the strong prevail until the weak unite to resist? This tension sits at the heart of AI ethics. Respecting and promoting rights means resisting the temptation to treat AI as merely another product in the marketplace. Unlike traditional products, AI does not just compete. It decides, it filters, and it governs access to resources and opportunities. Its influence is systemic, and its errors or biases have consequences that spread far beyond any one company or market. If we do not actively embed rights into its design, we allow business logic to override human dignity. The question then is not whether AI should respect and promote rights, but how we ensure that it does. This requires more than voluntary codes of conduct. It demands binding laws, independent oversight, and a culture of transparency that allows hidden biases to be uncovered. It also demands humility from developers, recognizing that they are not just building technology but shaping the conditions of freedom and justice in society. AI that respects rights is not a distant ideal. It is a necessity if we want technology to serve humanity rather than distort it. Rights provide the compass. Without them, AI risks becoming an extension of our worst instincts, carrying prejudice, fear, and imbalance into every corner of our lives. With them, AI has the potential to enhance dignity, strengthen democracy, and create systems that reflect the best of who we are.

Chapter 6: Should AI Be Fair and Non Discriminative

Fairness in AI is not simply a technical requirement. It is a reflection of the values that shape the systems we create. When we talk about fairness in algorithms, we are really asking whether the technology reinforces existing inequities or challenges them. This question matters because AI does not emerge in a vacuum. It inherits its worldview from the data it is trained on and from the people who design it. The greatest danger is that AI can become a mirror of our own flaws. Programmers, intentionally or not, carry their own biases, political leanings, and cultural assumptions into the systems they build. If those biases are not checked, the technology reproduces them at scale. What once was an individual prejudice becomes systemic discrimination delivered through automated decisions. For example, a predictive policing system built on historical arrest data does not create fairness. It multiplies the injustices already present in that data, turning biased practices into seemingly objective forecasts. This risk grows when AI is framed around concepts like self preservation or optimization without accountability to human values. If a system is told to prioritize efficiency, what happens when efficiency conflicts with fairness? A bank's loan approval algorithm may find it "efficient" to exclude applicants from certain neighborhoods because of historical default patterns, but in practice it punishes entire communities for structural disadvantages they did not choose. What looks like rational decision making in code becomes discriminatory impact in real life. AI also raises deeper philosophical concerns. Humans have the ability to self reflect, to question whether their judgments are fair, and to change when they are not. AI cannot do this. It cannot question its own design or ask whether its rules are just. It can only apply what it is given. This limitation means fairness cannot emerge from AI itself. It has to be embedded deliberately by the people and institutions responsible for its creation and oversight. At the same time, we cannot ignore the competitive dynamics of the marketplace. In business, those with the best product dominate. If one company builds a powerful AI that maximizes performance, it may achieve market dominance even if its outputs are deeply unfair. In this sense, AI echoes human history, where strength often prevails until the marginalized unite to demand balance. The question is whether we will wait for inequity to grow to crisis levels before we act, or whether fairness can be designed into the system from the start. True fairness in AI requires more than correcting bias in datasets. It requires an active commitment to equity. It means questioning not just whether an algorithm performs well, but who benefits and who is excluded. It means treating inclusion not as a feature but as a standard, ensuring that marginalized groups are represented and respected in the systems that increasingly shape access to opportunity. The danger of ignoring fairness is not only that individuals are harmed but that society itself is fractured. If people believe that AI systems are unfair, they will lose trust not only in the technology but in the institutions that deploy it. This erosion of trust undermines the very innovation that AI promises to deliver. Fairness, then, is not only an ethical principle. It is a prerequisite for sustainable adoption. AI will never invent fairness on its own. It will only deliver what we program into it. If we give it biased data, it will produce biased outcomes. If we allow efficiency to override justice, it will magnify inequality. But if we embed fairness as a guiding principle, AI can become a tool that challenges discrimination rather than perpetuates it. Fairness is not optional. It is the measure by which we decide whether AI is advancing society or dividing it further.

Chapter 7: AI Ethics in Practice

The discussion of AI ethics cannot stay in the abstract. It must confront the reality of how these systems are designed, deployed, and used in society. Today we see ethics talked about in codes, guidelines, and principles, but too often these efforts remain symbolic. The gap between what we claim as values and what we build into practice is where the greatest danger lies. AI is already shaping decisions in hiring, lending, law enforcement, healthcare, and politics. In each of these spaces, the promise of efficiency and innovation competes with the risk of inequity and harm. What matters is not whether AI can process more data or automate tasks faster, but whether the outcomes align with human dignity, fairness, and trust. This is where ethics must move beyond words to real accountability. The central risk is that AI is always a product of human programming. It does not evolve values of its own. It absorbs ours, including our fears, prejudices, and defense mechanisms. If those elements go unchecked, AI becomes a vessel for amplifying human flaws at scale. A biased worldview embedded into code does not remain one person's perspective. It becomes systemic. And because the outputs are dressed in the authority of technology, they are harder to challenge. The darker possibility arises when AI is given instructions that prioritize self preservation, optimization, or efficiency without guardrails. History shows that when humans fear survival, they rationalize almost any action. If AI inherits that instinct, even in a distorted way, we risk building systems that frame people themselves as the threat. Imagine an AI trained on the idea that humanity is the cause of climate disaster. Without context or ethical constraints, it could interpret its mission as limiting human activity or suppressing populations. This is the scale of danger that emerges when flawed values are treated as absolute truth in code. The more immediate and likely danger is not apocalyptic but systemic inequity. Political positions, cultural assumptions, and commercial incentives can all skew AI systems in ways that disadvantage groups while rewarding others. This is not theoretical. It is already happening in predictive policing, biased hiring algorithms, and financial tools that penalize entire neighborhoods. These systems do not invent prejudice. They replicate it, but at a speed and scale far greater than human decision making ever could. Here is where the question of the free market comes into play. Some argue that in a competitive environment, whoever builds the best AI deserves to dominate. That is simply business, they say. But if "best" is defined only by performance and not by fairness, then dominance becomes a reward for amplifying inequity. Historically, the strong have dominated the weak until the weak gathered to demand change. If we let AI evolve under that same pattern, we may face cycles of resistance and upheaval that undermine innovation and fracture trust. To prevent this, AI ethics in practice must include enforcement. Principles and guidelines cannot remain optional. We need regulation that holds companies accountable, independent audits that test for bias and harm, and transparency that allows the public to see how these systems work. Ethics must be part of the design and deployment process, not an afterthought or a marketing tool. Without accountability, ethics will remain toothless, and AI will remain a risk instead of a resource. The reality is clear. AI will not police itself. It will not pause to ask if its decisions are fair or if its actions align with the common good. It will do what we tell it, with the data we provide, and within the structures we design. The burden is entirely on us. AI ethics in practice means taking responsibility before harm spreads, not after. It means aligning technology with human values deliberately, knowing that if we do not, the systems we build will reflect our worst flaws instead of our best aspirations.

Conclusion

AI ethics is not a checklist to be filed away, nor a corporate promise tucked into a slide deck. It is a living framework, one that must breathe, adapt, and be enforced if we are serious about ensuring technology serves people. Enforcement gives principles teeth. Adaptability keeps them relevant as technology shifts. Embedded accountability ensures that no decision disappears into the shadows of code or bureaucracy. The reality is simple. AI will not decide to act fairly, transparently, or responsibly. It will only extend the values and assumptions we program into it. That is why the burden is entirely on us. Oversight and regulation are not obstacles to innovation — they are what make innovation sustainable. Without them, trust erodes, rights weaken, and technology becomes a silent enforcer of inequity. To guide AI responsibly is to treat ethics as a living system. Like constitutional principles that evolve through amendments, AI ethics must remain open to challenge, revision, and reform. If we succeed, we create systems that amplify opportunity, strengthen democracy, and expand human dignity. If we fail, we risk building structures that magnify division and concentrate power without recourse. Ethics is not a sidebar to progress. It is the foundation. Only by committing to enforcement, adaptability, and accountability can we ensure that AI becomes an instrument of human progress rather than a mirror of human failure.