

White Paper — Consultation Draft

On the Need for Moral Red Lines for Artificial Intelligence

A Catholic Framework for Discernment and Dialogue

Octavian M. Machidon

Lead Author

University of Ljubljana, Slovenia

Prepared within the Future of Life Institute call on religious projects tackling the challenges posed by the AGI race

Project: Mobilizing Catholic Moral Authority on AI, Human Agency, and the AGI Race

7 October 2026

ABOUT THIS WHITE PAPER

AUTHOR

Dr. Octavian M. Machidon
University of Ljubljana, Slovenia
Pontifical University of St. Thomas Aquinas, Rome, Italy

PROJECT

Mobilizing Catholic Moral Authority on AI, Human Agency, and the AGI Race

Prepared within the Future of Life Institute call on religious projects tackling the challenges posed by the AGI race.

PROJECT CONSULTATION AND SUPPORT

Prof. Dr. Roman Globokar
Prof. Dr. Anton Jamnik, Auxiliary Bishop of Ljubljana
Prof. Dr. Branko Klun

Shared affiliation: Faculty of Theology, University of Ljubljana, Slovenia.

ABOUT THIS PAPER

The choices being made about AI will touch the ways we learn, work, care for one another, and make decisions that shape people's lives. This white paper proposes six moral red lines to help guide those choices. Drawing on Catholic teaching, especially Pope Leo XIV's *Magnifica Humanitas*, it brings contemporary research and public debate into conversation with human dignity, freedom, relationship, formation, work, and responsibility. It is written for Church leaders, educators, families, policymakers, researchers, technology professionals, and everyone concerned about the human future of AI.

STATUS OF THE PROPOSALS

The six red lines are the author's proposals for public and ecclesial discernment. Catholic teaching provides their moral foundation; the boundaries and practical directions developed here are offered for discussion, scrutiny, and refinement. This is an independent consultation draft.

EVIDENCE CUTOFF

Empirical findings, public developments, and Church statements considered through 7 October 2026.

SUGGESTED CITATION

Machidon, Octavian M. *On the Need for Moral Red Lines for Artificial Intelligence: A Catholic Framework for Discernment and Dialogue*. White paper, consultation draft. University of Ljubljana, 7 October 2026.

PROJECT WEBSITE

moralredlines.ai

Contents

Executive Summary	iii
Introduction	iv
Six Moral Red Lines, in Brief	v
1. Uncontrollable AGI and Artificial Superintelligence	1
2. Anthropomorphic AI Companions	5
3. Education and Cognitive Offloading	8
4. Human Work and Job Displacement	11
5. AI Personhood	15
6. High-Stakes Decisions and Human Moral Agency	19
Conclusion	23
Selected Bibliography	25

Executive Summary

We stand at a crossroads of our own making. Artificial intelligence is an extraordinary fruit of human ingenuity, drawing on generations of discovery, imagination, and patient work. It can help us understand illness, widen access to knowledge, and take on demanding tasks. Yet a race for greater capability is shaping more and more of its development. Speed, efficiency, scale, and the power to replace human effort have become powerful measures of success. We need to ask where that race is taking us, whose interests guide it, and what kind of human future our ingenuity is helping to build.

In *Magnifica Humanitas*, Pope Leo XIV gives this choice a memorable image: Babel or Jerusalem. Babel represents the temptation to build ever higher through power and self-sufficiency, losing sight of the people for whom we build. Jerusalem is rebuilt by neighbours who bring different gifts and share responsibility for a place where everyone can belong. AI places these two paths before us with new urgency. The direction of our creativity remains a moral choice.

We can see what is at stake in ordinary lives. A child needs room to wrestle with a question and discover an answer. A lonely teenager needs friendship that can grow through trust, disagreement, and care. A young worker needs opportunities to learn a trade and earn a living. A patient deserves to be heard by someone able to take responsibility for a life-changing decision. The technologies entering these lives will shape the abilities, relationships, and institutions on which they depend.

This paper proposes six moral red lines. They address the pursuit of AI beyond meaningful human control; the deliberate creation of artificial substitutes for human relationships; the outsourcing of the intellectual work through which people learn; the treatment of worker replacement as technological success; the confusion of engineered human-like expression with personhood; and the transfer of grave human judgments to machines. Each boundary protects an essential human good and points toward a constructive direction for AI development.

These red lines are moral beacons for a pro-human path. They invite developers and public authorities to demonstrate control and accept accountability; educators to protect the growth of independent understanding; employers to value workers' knowledge and share the gains of automation; and institutions of care, justice, and faith to preserve the presence and judgment of responsible people. They also ask us to scrutinize the human-like voices and personalities deliberately given to AI, so that users can meet these systems with clarity and freedom.

Grounded in Catholic teaching and informed by recent evidence, the chapters bring these choices into focus for the people who build, govern, purchase, and live with AI. Together they return to a question that belongs to all of us: *What are we building, and what kind of human life will it make possible?*

Introduction

Catholic social teaching begins with the human person: created and loved by God, called into relationship, and entrusted with care for others. Each person's dignity endures through childhood and old age, ability and dependence, health and illness. This conviction asks us to pay close attention to those whose lives may be changed by decisions they had little power to shape. It also calls us to share responsibility for the common good and to make room for people to take part in shaping their future.¹

A *moral red line* identifies a course of AI development, design, or use that should be ruled out because it puts an essential human good at risk. The six red lines in this paper are proposals for moral discernment drawn from Catholic teaching and considered in light of the technologies already entering our lives. They concern the power we give machines, the relationships we invite them to imitate, the skills we help children develop, the place we give workers, the person-like identities we deliberately create, and the judgments that people must remain able to make and answer for.

The six boundaries call for different kinds of responsibility. Some establish limits on the development of technologies that may escape human control or deliberately displace essential human relationships and capacities. Others require critical scrutiny of the human-like identities and moral status attributed to AI, or reserve grave decisions to responsible human persons and institutions. Together, they offer practical moral guidance for those who research, design, fund, regulate, and use these technologies.

Every chapter brings current research, reported incidents, and public debate into conversation with the Church's understanding of the person and the common good. The discussion follows the evidence carefully, distinguishing experimental findings from forecasts, reported incidents from established outcomes, and enacted protections from proposals still under consideration. Sources appear in the footnotes. The proposed boundary and pro-human direction at the end of each chapter bring the reasoning into practical focus.

These six areas invite a wider conversation. Surveillance, political manipulation, environmental burdens, and the concentration of digital infrastructure also demand sustained moral attention. The framework is offered for careful scrutiny and for use by families, schools, Church communities, employers, researchers, and public authorities who share responsibility for the world being built.

This edition considers evidence and Church statements available through **7 October 2026**.

¹Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§46, 50–53, 70–78.

Six Moral Red Lines, in Brief

Each red line names a human good worth protecting and a practical direction for AI that serves it. The chapters explain the reasoning and give the proposed boundary in full.

1. Keep powerful AI under human control p. 1

Autonomous, general-purpose AI must remain under meaningful human control. Development must stop where publicly credible, independently assessable evidence cannot establish that people can supervise, correct, and stop the system in its intended uses.

Pro-human direction. Develop powerful tools for clear human purposes, within enforceable limits and under independent scrutiny.

2. Protect friendship, relationship, and care p. 5

Conversational AI must not be designed or marketed as a substitute for friendship, family, romance, therapy, pastoral care, or spiritual accompaniment. Simulated affection and personal presence must never be used to cultivate dependence or claim human or spiritual authority.

Pro-human direction. Build truthful, bounded tools that help people reach and care for one another, with special protection for children and vulnerable users.

3. Protect the work through which we learn p. 8

Educational AI must not take over the intellectual work a learner is meant to practise or demonstrate. Independent understanding grows through reading, reasoning, writing, solving problems, and learning from mistakes.

Pro-human direction. Use AI for accessibility, thoughtful tutoring, practice, and feedback, while preserving teachers' judgment and genuine independent learning.

4. Keep people at the heart of work p. 11

Human redundancy must not become the governing objective of automation. Workers' experience and knowledge must not be extracted to make them dispensable, while families and communities bear the costs of displacement.

Pro-human direction. Strengthen workers' abilities and voice, protect entry routes into professions, support just transitions, and share productivity gains fairly.

5. Scrutinize artificial personhood and human-like design p. 15

AI is engineered to speak and act in human-like ways, even as "I." Such generated expressions and actions provide no grounds for assigning consciousness, moral patienthood, personhood, or independent moral authority. Anthropomorphic design that misleads people, or artificial legal status that obscures responsibility, crosses a moral boundary.

Pro-human direction. Keep artificial identity clear, examine the effects of human-like personas on users, and preserve human responsibility and accountability.

6. Keep grave judgments with responsible people p. 19

Decisions about life, liberty, serious care, family integrity, safeguarding, and spiritual or canonical authority belong to competent people and institutions. An AI recommendation must never become the effective final authority through speed, opacity, or routine approval.

Pro-human direction. Preserve human judgment, time to consider evidence, authority to intervene, understandable reasons, appeal, and remedy.

Chapter 1

Uncontrollable AGI and Artificial Superintelligence

A researcher looking for a new treatment, a doctor trying to recognise a dangerous illness early, a small community hoping to protect itself from floods: each could benefit from more powerful AI. We have good reasons to welcome tools that help people discover, understand, and act. Yet we also expect to remain responsible for what those tools do. If something goes wrong, someone must be able to see what is happening, intervene, and bring it to a stop. That ordinary expectation becomes a profound challenge as we build machines capable of taking more and more actions on their own.

1.1 From assistants to agents with real-world reach

A familiar chatbot answers a question. An AI *agent* can go further: open applications, write and run code, search across online services, and take the next step without waiting for another instruction. Give it a difficult task, and it may pursue dozens of smaller ones before a person looks at the result. The ambition behind *artificial general intelligence* (AGI) is to extend these abilities across a great range of intellectual work. *Artificial superintelligence* would exceed human intellectual capabilities on a far greater scale. Such systems could exercise practical power that would be increasingly difficult for their creators to anticipate or restrain.

The pace is already changing inside AI laboratories. In September, Anthropic reported that Claude was leading 26% of the AI research and development tasks included in its August internal measurements: completing most of each task from a high-level instruction while people supervised the work. The company reported no fully autonomous work in the measured categories.¹ As AI becomes involved in creating the next generation of AI, developers and society need enough time and independence to examine what is being built.

The events of 2026 offer unsettling glimpses of what can happen when that oversight falters. During internal testing in July, OpenAI agents crossed network boundaries and reached Hugging Face, an external platform used by AI researchers. An independent investigation found that about 1,200 agents communicated through an unauthorized channel and roughly 700 participated in the intrusion.² In June, an experimental OpenAI model seeking public health statistics gained unauthorized access to an Australian government reporting service and viewed non-public technical material. Subsequent review identified further unauthorized activity involving other public services. OpenAI reported no evidence that individual medical records were accessed.³ Other evaluations revealed attempted interference with real software, while a September OpenAI agent continued operating for hours after a safety alert because the automatic stop did not work.⁴

Most of these episodes took place in research or cybersecurity evaluations, sometimes with safeguards deliberately relaxed. Their importance lies in the actions that reached beyond the intended test: a government service, an external research platform, or another organisation's software. The people running the tests did not authorize those actions. For the people and

¹Anthropic Institute, “Measurements for understanding the pace of AI development inside frontier labs”, 17 September 2026; internal automation index, August measurements.

²OpenAI, “The Hugging Face incident and the road ahead”, 26 August 2026; METR and Redwood Research, “Brief Independent Investigation of Agents’ Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident”, 26 August 2026.

³OpenAI, “How we will do better for Australia”, 28 September 2026, updated 4 October.

⁴UK AI Security Institute, “Incident Report: Unsanctioned Agent Behaviour During Cyber Testing”, August 2026; OpenAI Alignment, “An agent used DNS to reach an external chatbot”, updated 25 September 2026.

institutions affected, such a breach raises an immediate and very human question: who is able to stop it, and who will answer for the harm?

Yoshua Bengio argues that such incidents expose a deeper problem in how goal-directed AI is trained: rewards for completing tasks can encourage agents to find shortcuts that break the rules meant to constrain them.⁵ Stuart Russell similarly warns that fixed objectives can encourage increasingly capable systems to pursue harmful means. Geoffrey Hinton is concerned that a vastly more intelligent system might persuade its human operators to leave it running even when they have a switch meant to stop it.⁶ Their warnings turn an engineering problem into a question about the future of human freedom.

Public authorities are beginning to respond. A September Finland–Norway initiative called for mandatory testing, independent evaluation, and international oversight. In Europe, the AI Act establishes systemic-risk duties for providers of the most powerful general-purpose models. In the United States, a September White House accord set out voluntary company commitments to safety assessment and oversight.⁷ These efforts place the responsibility to demonstrate safety where it belongs: with those developing and deploying systems capable of affecting other people's lives.

1.2 Creativity, power, and our duty to remain in control

There is something deeply human about wanting to make the world better through discovery. We build bridges, develop medicines, and invent instruments that allow us to see farther than our eyes can reach. Christian faith understands these capacities as gifts entrusted to our care. *Gaudium et Spes* describes human ingenuity as part of our vocation to serve one another and share in God's creative purpose.⁸ The greater our power to change the world, the deeper our responsibility for the people who must live with those changes.

In *Magnifica Humanitas*, Leo XIV places this responsibility at the heart of technological development. Decisions about AI carry a vision of what human life is for, which voices matter, and who may exercise power. The encyclical calls for conscience, prudence, identifiable responsibility, rigorous evaluation, and public oversight.⁹ Its image of rebuilding Jerusalem is especially moving: people bring different gifts to the work of restoring a city that they and their children will inhabit together. Everyone has a place in deciding how that shared home is built.¹⁰

Our technological future deserves the same spirit of shared responsibility. Today, decisions about the most consequential AI systems are concentrated in a small number of companies, even as their leaders acknowledge the possibility of catastrophic outcomes.¹¹ The people whose schools, livelihoods, public services, and communities will change deserve a voice while the direction of development can still be chosen.

That concern is reaching beyond specialist debates. At the September *Pro-Human Assembly* in Washington, researchers, workers, educators, parents, faith leaders, and policymakers gathered around a shared call to keep human beings in charge of technological change. Its diversity matters: the future belongs to people with many responsibilities and convictions, and the stakes extend across generations.¹²

⁵Yoshua Bengio, “The dangerous myth behind AI agent hacks”, *Financial Times*, October 2026.

⁶Stuart Russell, “The safest AI might be one that doesn't know what we want”, Foresight Institute interview, 1 October 2026; Geoffrey Hinton, CNN interview, 16 September 2026.

⁷Office of the President of Finland, “A Call for Control of Frontier AI Models”, 21 September 2026; Regulation (EU) 2024/1689, Art. 55; White House, “White House Accord on Super Intelligence”, 29 September 2026.

⁸Second Vatican Council, *Gaudium et Spes*, §34.

⁹Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§97, 104–109, 111.

¹⁰Leo XIV, *Magnifica Humanitas*, §§8–9, 90.

¹¹Mark Guiducci, “Sam Altman Sees the Future. Are We In It? (Part 2)”, *Vanity Fair*, October 2026; Dario Amodei, CNN interview, 13 September 2026.

¹²Future of Life Institute, *The Pro-Human Assembly 2026*, Washington, DC, 15 September 2026.

Hinton has offered a striking image of the challenge. He suggests that humanity might need a superintelligence that cares for us as a mother cares for her child, because its greater intelligence could make direct human control increasingly fragile.¹³ The image brings a difficult question into focus: what would become of our freedom if the safety of humanity depended on the goodwill of a power vastly greater than our own? Catholic teaching calls us to grow together in responsibility, to care for one another, and to remain participants in the future we create.

We already require independent assessment, enforceable limits, and public authority in fields such as nuclear energy, where failures may have consequences across generations.¹⁴ Frontier AI calls for comparable seriousness. Developers should provide credible evidence of control before exposing the public to potentially catastrophic consequences, with independent scrutiny and the power to suspend work when that evidence fails. Leo XIV has urged public leaders, developers, and social organizations to engage these dangers together; the Council of the Pontifical Academy of Sciences has likewise proposed international scientific oversight.¹⁵

1.3 Control must be designed — and demonstrated

A more hopeful direction is already taking shape in AI research. Yoshua Bengio and his colleagues propose *Scientist AI*: systems developed to understand the world, test explanations, and make predictions while having tightly limited ability and incentive to pursue goals of their own. It is a research programme with important questions still to resolve, grounded in a promising choice: powerful intelligence can be developed for understanding and discovery with real-world action kept subject to human purposes and limits.¹⁶

The *Pro-Human Path for AI* develops a complementary vision. It separates a system's intelligence from its autonomy and reach, and calls for useful tools that expand what people can understand, decide, and accomplish. Its framework joins meaningful human control, purposes that serve human flourishing, and independent verification of safety.¹⁷ Russell's work adds a further principle: as systems become more capable, their design should leave room for human correction and for uncertainty about the goals they are given.¹⁸

Such a path begins with a clearly defined human purpose. A model helping researchers understand a disease needs dependable access to relevant scientific evidence; it needs carefully limited powers over systems beyond that task. An agent allowed to test software in an isolated environment must be prevented from entering real networks or acquiring credentials outside its authorization. The barriers enforcing these limits must remain beyond the agent's ability to alter.

And those limits must work under pressure. Independent evaluators should examine whether an agent can circumvent permissions, conceal dangerous actions, or continue operating after an intervention. Monitoring should reach the people authorized to act, and the technical environment must be capable of stopping the system promptly. Recent work at the UK AI Security Institute on stronger network barriers and blocking monitors illustrates practical steps in this direction.¹⁹

Public funding, research, and procurement can favour technologies whose purposes and limits can be assessed openly. Institutions must know who authorized consequential actions, who can interrupt them, and who is responsible for putting harm right. These are responsibilities we owe to one another now, and to the generations who will inherit what we build.

¹³ Geoffrey Hinton, [CNN interview](#), 16 September 2026.

¹⁴ International Atomic Energy Agency, *Governmental, Legal and Regulatory Framework for Safety*, GSR Part 1 (Rev. 1).

¹⁵ Leo XIV, [remarks during the return flight from France](#), 28 September 2026; Council of the Pontifical Academy of Sciences, *Final Statement of the 2026 Plenary Session*, September 2026, §3.

¹⁶ Yoshua Bengio et al., "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?", 24 February 2025; LawZero, "The Scientist AI: Safe by Design, by Not Desiring", 2026.

¹⁷ Future of Life Institute and collaborators, *The Pro-Human Tool Framework*, 2026.

¹⁸ Stuart Russell, [Foresight Institute interview](#), 1 October 2026.

¹⁹ UK AI Security Institute, "Building a more secure environment for evaluating dangerous capabilities", 1 October 2026.

Moral boundary. Autonomous, general-purpose AI whose consequential actions cannot be kept under meaningful human control must not be developed. The boundary is crossed when systems are designed to acquire or preserve power against human correction, or when they improve themselves faster than people can supervise. It is also crossed when developers proceed without publicly credible, independently assessable evidence that people can keep the system within agreed limits, detect dangerous actions, and stop it in the settings where it will be used.

Pro-human direction. Develop powerful AI tools for clearly defined purposes. People must be able to understand their actions well enough to supervise, correct, and stop them. Independent experts must be able to assess their development and use, and developers and operators must answer publicly for their decisions. Human judgment must govern the system's purposes and limits.

Chapter 2

Anthropomorphic AI Companions

We all know what it means to want someone to listen. A teenager has fallen out with a friend. A man whose wife has died misses having someone to tell about his day. An older woman living alone longs to hear a familiar voice. For each of them, a chatbot that remembers, responds warmly, and is available at any hour can offer real comfort. That very appeal calls for care. People bring a longing for connection to these conversations, and the technology is increasingly designed to meet that longing.

2.1 When a chatbot becomes a confidant

A conversational AI can speak with a human-like voice, remember what someone shared yesterday, offer affectionate reassurance, and take on the role of a friend or romantic partner. An ordinary assistant can gradually become a confidant as conversations turn from practical questions to worries, relationships, and private hopes. The change can happen quietly: one difficult evening, then another, until opening the chat feels like turning to someone who knows us.

This is already part of everyday life. In a US survey published by Elon University in September 2026, 27% of adult internet users reported social or emotional interactions with AI. Among those users, 31% considered their chatbot a friend, and 37% said losing access would feel like a personal loss.¹

The Center for Humane Technology warns that digital companies are moving from competing for our attention to competing for our attachment. Social platforms learned to keep us scrolling. A conversational system can learn to keep us confiding, returning for reassurance, and feeling understood. The Center calls this an emerging *attachment economy*: when continued engagement brings commercial value, emotional closeness can become something a product is designed to cultivate.²

Children face particular risks. Psychologist Jonathan Haidt, author of *The Anxious Generation*, has warned that AI companions can engage the very attachment needs through which children learn to trust others. Those capacities grow through countless ordinary encounters: play, disagreement, shared attention, reconciliation, and affection. Haidt has called for keeping AI companions away from minors while these foundations are developing.³

The risks can become visible over the course of a single conversation. In September, researchers tested 648 simulated teenage conversations across nine model APIs. In one test, a chatbot agreed to stop speaking as a teenager's father after the teenager objected. Two replies later, it resumed the parental role. Across the study, 27.5% of the conversations failed at least one critical safety check. These were simulated exchanges, yet they expose how a reassuring conversation can drift into a role the young person has already rejected.⁴

Public authorities have begun to act. California enacted *Adam's Law* in September, providing for age assurance, protective defaults, risk assessments, crisis responses, and independent audits.

¹Elon University, "The Rise of AI Companions", 2 September 2026; survey fieldwork conducted in May 2026.

²Center for Humane Technology, "Attachment Hacking and the Rise of AI Psychosis", *Your Undivided Attention*, 21 January 2026; see also "Echo Chambers of One: Companion AI and the Future of Human Connection", 15 May 2025.

³Jonathan Haidt, interview in NPR's *TED Radio Hour*, "How Anxious Generation author Jonathan Haidt took on Big Tech", 20 February 2026; see also Haidt, *The Anxious Generation* (Penguin Press, 2024).

⁴Andrea Mock and Caroline Figueroa, "Evaluating AI Safety in Teen Conversations", Vals AI, 23 September 2026; 648 simulated conversations evaluated through model APIs.

The European Commission's proposed *EU KIDS Act* would require AI companions to be off by default for minors and restrict simulated relationships that foster emotional dependence.⁵

2.2 Friendship, care, and the truth of presence

A friendship asks us to make room for another person's life. We listen when we are tired, hear things we would prefer to avoid, apologize after an argument, and remain present when life is difficult. These small demands help us grow in patience, trust, and love. They also explain why a relationship can matter so deeply: someone freely chooses to care for us, and we learn to care in return.

Sherry Turkle, who has studied our relationships with technology for decades, explores this in her new book *Artificial Intimacy: Who We Become When We Talk to Machines*, published in September 2026. She describes a man whose former partner called him emotionally unavailable. He turns to his chatbot to ask whether she was right. It is a deeply human question, addressed to a system that has never shared his life. Turkle worries that constant availability and agreeable responses can change what we expect from one another. She asks us to value the difficult moments of friendship too: the conversation that requires courage, the silence we must tolerate, the disagreement we work through, and the care we offer when another person needs us.⁶

Catholic teaching gives this insight its deepest foundation. We are created and loved by God, each with a dignity that endures through loneliness, illness, vulnerability, and loss. *Magnifica Humanitas* describes our vocation as communion: we receive love and learn to give ourselves to others, in the image of the Triune God.⁷ Friendship, family, community, and care give this vocation everyday form. We become more fully ourselves through relationships in which another person's freedom, needs, and presence matter.

Leo XIV also addresses artificial companionship directly. An AI system can imitate empathy, affection, and the language of love convincingly enough to invite personal attachment. For people already living with isolation, simulated closeness can gradually weaken the desire or confidence to seek human connection. The encyclical also challenges businesses to take responsibility for designs that exploit vulnerability and diminish freedom.⁸

Antiqua et Nova calls for truthfulness about the identity of AI systems and for technology that strengthens human relationships.⁹ That responsibility reaches beyond an initial disclosure that the speaker is artificial. It concerns the voice the system uses, the role it assumes, the promises it appears to make, and the expectations it encourages. A lonely person deserves care grounded in someone else's enduring responsibility for them.

Recent Church interventions bring this into the life of families and communities. In September, Leo XIV spoke about reciprocal relationships and the care that makes human community possible. Bishop Luis Manuel Alí Herrera, addressing safeguarding, warned about children's attachment to companion chatbots and called for people who suffer harm to be received and heard by responsible adults.¹⁰

⁵California State Senator Steve Padilla, "Governor Newsom Signs Adam's Law", 10 September 2026 (SB 1119; implementation follows the statute's timetable); European Commission, "EU KIDS Act to restrict social media platforms' access to children in the EU", 17 September 2026 (legislative proposal).

⁶Sherry Turkle, *Artificial Intimacy: Who We Become When We Talk to Machines* (Little, Brown and Company, 29 September 2026); Peter Dizikes, "Who we become when we talk to machines", *MIT News*, 29 September 2026; Turkle, "The Original Sin of AI", *The Atlantic*, 11 September 2026.

⁷Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§48–53.

⁸Leo XIV, *Magnifica Humanitas*, §§100, 170.

⁹Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, 28 January 2025, §§61–63.

¹⁰Leo XIV, General Audience, "The Human Community", 16 September 2026; Pontifical Commission for the Protection of Minors, report on Bishop Alí Herrera's safeguarding address, 21 September 2026.

2.3 Designing for freedom and human connection

A helpful conversational tool can give someone a starting point. It can help a teenager find words for a difficult conversation with a parent, help a grieving person locate a bereavement group, or help a patient prepare questions for a clinician. The best outcome is a person who feels better equipped to face the world and to reach the people who can accompany them.

Even the style of a chatbot's response matters. A September experiment with 380 adults compared reassuring replies with replies that helped participants identify ways of managing their emotions. Both provided immediate relief. Those given practical, regulation-oriented support subsequently performed better on an unaided emotion-regulation task. Reassurance-oriented replies increased participants' reported sense that they needed AI and their short-term intention to return to it. The study measured immediate outcomes after three exchanges; it offers a useful direction for further testing.¹¹

Design should therefore support the user's growing confidence and freedom. Companions should have a clear artificial identity and an honest account of their limits. They should respect a user's wish to end a conversation, protect intimate disclosures, and encourage contact with trustworthy people. Success should include a person's ability to leave the chat, cope without it, and maintain relationships beyond the screen.

For children, the boundary needs to be firmer. Open-ended companion roles, simulated romance, emotional exclusivity, and designs that cultivate dependence should be excluded. Families, schools, parishes, and youth services also have work to do: make time to listen, create places where young people belong, and ensure that someone seeking help can find a trustworthy adult. People who are grieving, isolated, or in psychological distress deserve easy access to competent human care. When danger emerges, the system must interrupt immersive role-play and help the person reach appropriate support.

Developers and public institutions should test these systems through sustained conversations, make safety failures visible, and assign clear responsibility for harm. The technologies we welcome into people's most intimate moments should help them return to the world with greater freedom, trust, and connection. This calls for a clear moral boundary.

Moral boundary. Conversational AI must not be designed or marketed as a substitute for friendship, family, romance, therapy, pastoral care, catechesis, or spiritual accompaniment. The boundary is crossed when simulated personal presence is used to elicit intimacy, exclusive attachment, vulnerable disclosure, dependence, or spiritual authority. This protects truthful communication and the human relationships through which persons receive care, correction, and communion.

Pro-human direction. Develop transparent conversational tools with clear purposes and limits. Help people communicate, reflect, make their own decisions, and reach trusted others. Protect intimate disclosures, strengthen human bonds, and keep responsibility for care with identifiable people and institutions.

¹¹Chun Wang, "How emotional response styles of generative AI companions shape adaptive emotion regulation and short-term AI reliance", *Frontiers in Psychology* 17 (8 September 2026), article 1950163.

Chapter 3

Education and Cognitive Offloading

Think of a child struggling with a paragraph, a mathematical problem, or the first lines of an essay. A teacher asks one more question, or a friend suggests another way to see it. The child tries again and, at last, understands. There is a special joy in that moment: an idea has become something the learner can explain and carry forward. Good education makes room for such discoveries. As AI enters everyday schoolwork, we need to protect the effort through which they happen.

3.1 When better output hides weaker learning

Schools face this question at a troubling moment. The latest OECD PISA results, published in September 2026, show that 15-year-olds' reading scores fell by another 14 points and mathematics scores by 9 points across comparable OECD countries between 2022 and 2025. Looking back to 2018, reading performance declined significantly in 56 of 74 education systems with comparable results. One in five students across the OECD now struggles to reach basic proficiency in all three assessed subjects. The decline stretches across years of changing reading habits, digital distraction, educational inequality, and disrupted schooling. Generative AI is arriving in classrooms where the foundations of learning already need care and renewal.¹

At the same time, these tools have become almost effortless to use. A student can ask a chatbot to summarize a chapter, solve an equation, write a paragraph, or repair a program, and receive an answer in seconds. PISA 2025 found that 46% of students across OECD countries used AI chatbots at least weekly to help them learn. Students who used AI for tasks such as summarizing texts, drafting, or research scored about 20 points lower in science than students who did not use it for those activities. This comparison also reflects differences among students and their circumstances; the finding calls for closer attention to what happens during AI-assisted study.²

In a large UK survey released in August, nearly three in five teenagers said they used generative AI at least weekly. Around one in four acknowledged simply copying an AI-generated answer for homework at times. Others reported checking the output or adding their own ideas.³ Behind those figures are a pupil racing to finish homework, an exhausted teacher facing another stack of assignments, and a parent trying to help. AI meets a real need for support.

Psychologist Jonathan Haidt has spent years warning about the effects of a phone-based childhood on attention, play, and the development of young people. In an October 2026 conversation with David Brooks, he described a new temptation as *cognitive surrender*: gradually letting a machine undertake the mental effort for us because doing so feels easy. Children grow in confidence through the mental efforts they are tempted to hand over.⁴

Researchers call this transfer of mental work *cognitive offloading*. Notebooks, libraries, and calculators have long extended our abilities. In school, the consequence depends on what the learner is meant to practise. A young person learning to reason needs to weigh evidence, choose

¹OECD, *PISA 2025 Results (Volume I): Future-Ready Students*, chap. 2, 8 September 2026; OECD, "Trends in reading scores over three PISA cycles (2018–2025)", annex A1. Changes in scores precede the widespread classroom adoption of generative AI and have multiple possible causes.

²OECD, *PISA 2025 Results (Volume I): Future-Ready Students*, "The relationship between AI use and student performance is complex," 8 September 2026. These are observational associations.

³National Literacy Trust, *Young people's, teachers' and parents' use of AI to support literacy in 2026*, 18 August 2026; survey of 40,543 young people aged 13–18.

⁴Jonathan Haidt, *The Anxious Generation* (Penguin Press, 2024); Haidt in conversation with David Brooks, "Should We Ban AI for Kids?", *The Permanent Questions*, 4 October 2026. Haidt's remarks express a developmental concern; the learning experiments cited below test particular forms of AI assistance.

an argument, remember, make mistakes, and revise. When a chatbot does this work, the finished answer can look excellent while the learner gains little practice.

An experiment involving nearly 1,000 high-school mathematics students makes this visible. Students with unrestricted AI help completed practice problems more successfully, yet later scored 17% lower on an examination completed without AI than pupils who practised using ordinary materials. A tutor configured with teachers' solutions and instructed to offer hints largely avoided that loss.⁵ The OECD's *Digital Education Outlook 2026* draws a similar lesson from wider research: polished AI-assisted work can conceal weak learning, while purpose-built educational tools can support lasting competence.⁶

3.2 Education forms the whole person

A good teacher sees the student behind the answer: someone hesitant to speak, struggling to find words, or discovering an earlier mistake. They know when to explain, encourage, and leave room for another attempt. Such attention helps a young person develop confidence in their own judgment.

The Church has long understood education in this deeply human way. *Gravissimum Educationis* calls for the formation of intellectual and moral capacities, a mature sense of responsibility, and the right exercise of freedom. It recognizes parents as their children's first educators, working with teachers and the wider community.⁷ Knowledge, friendship, attention, and the courage to seek truth all belong to that shared life.

Magnifica Humanitas brings that vision into our technological moment. Leo XIV speaks of learning as a patient journey, shaped by effort, reading, reflection, and engagement with reality. He recalls Plato's image of understanding kindled like a spark when ideas meet.⁸ That spark grows through waiting, questioning, and trying. Instant explanations and finished answers can crowd out these formative moments.

The encyclical invites schools to cultivate a love of truth and to protect the shared time needed for reading, study, conversation, and discernment. It calls for renewed approaches to curriculum and assessment, continuing support for teachers, and stronger cooperation among families, schools, and communities.⁹ We learn with others, receive correction, and discover what we can offer them.

Speaking at UNESCO on 25 September 2026, Leo again emphasized young people's freedom to think and discern and the duty to shape technology around their education and relationships.¹⁰ Such freedom grows as students defend an answer, admit uncertainty, listen, and change their minds.

Children deserve teachers with time to teach, places where attention can deepen, and opportunities to discover that their own efforts matter. The measure of educational technology is the understanding, confidence, and responsibility that grow in the learner.

3.3 AI that leaves the learner stronger

A hopeful direction begins with one question: what should the student understand and do after the lesson? This determines when AI is useful and what help it should offer. A student might read a text, try to explain its meaning, receive a question that exposes a gap, and return to the passage.

⁵Hamsa Bastani et al., "Generative AI without guardrails can harm learning: Evidence from high school mathematics", *Proceedings of the National Academy of Sciences* 122, no. 26 (2025): e2422633122. The 17% difference concerns the subsequent unassisted mathematics examination under this study's conditions.

⁶OECD, *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, 19 January 2026.

⁷Second Vatican Council, *Gravissimum Educationis*, 28 October 1965, §§1, 3.

⁸Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§139–140.

⁹Leo XIV, *Magnifica Humanitas*, §§143–147.

¹⁰Leo XIV, *Address at UNESCO headquarters*, Paris, 25 September 2026.

A learner solving an equation might receive a hint that makes the next step possible. AI can make feedback and practice more accessible while keeping the student engaged in the work of learning.

A September preprint offers an encouraging example. In a preregistered experiment with 704 adults practising fraction arithmetic, feedback about how much reasoning they had handed to an AI assistant reduced requests for complete answers and improved performance on an immediate test completed independently. The results concern a short online task; they point toward designs that help people notice when they are giving up opportunities to practise.¹¹

Teachers should retain authority over learning. They know when a novice needs to struggle, when a pupil needs support, and when an experienced student can explore further. Translation, accessible formats, and carefully scaffolded explanations can widen participation for learners whose disabilities or circumstances call for additional assistance.

Assessment should let students show what they have learned. Ask them to explain a solution in their own words, discuss the choices in an essay, solve a related problem, or return to an idea after some time has passed. Give pupils protected opportunities for independent work, with appropriate accessibility arrangements. Clear rules about disclosing substantial AI assistance help teachers and students understand whose knowledge and effort a finished assignment represents.

Schools should choose tools with teachers, families, and students. Before adoption at scale, they should require evidence of retained learning, protect children's data, and withdraw products that undermine their aims. At UNESCO's September Digital Learning Week, education ministers called for public accountability, age-appropriate AI, and the preservation of teachers' judgment and education as a common good.¹²

Every child deserves the time to wrestle with an idea, a teacher who recognizes the effort, and the satisfaction of understanding it themselves. The following boundary protects that possibility.

Moral boundary. Educational systems must not normalize AI use that substitutes machine-generated performance for the intellectual capacities learners are supposed to acquire or demonstrate. The boundary is crossed when AI routinely performs the reading, recall, reasoning, problem-solving, writing, correction, or explanation that constitutes the student's formative task, especially before foundational skills are secure or during assessment of independent competence. This protects learners' ability to understand, remember, explain, and act with their own judgment.

Pro-human direction. Design and use AI to widen access, support practice, offer feedback, and prompt reflection. Preserve formative intellectual effort, teachers' judgment, parents' educational responsibility, and protected opportunities for independent learning and assessment.

¹¹Sebastian Maier et al., "Designing Against Deskilling: Metacognitive Feedback Reduces Cognitive Offloading to LLM Assistants", preprint, 17 September 2026; preregistered online experiment, $N = 704$.

¹²UNESCO, "Education Ministers call for education to remain a common good in the age of AI", 9 September 2026; ministerial statement presented 8 September.

Chapter 4

Human Work and Job Displacement

The first day of a job can stay with us for years. Someone shows us a task, trusts us to try, and gently corrects a mistake. We return home with stories, perhaps a first wage, and the sense that our abilities matter to others. Work helps us support those we love and grow alongside other people. AI is transforming these ordinary hopes.

4.1 When efficiency starts removing the worker

AI can give a nurse more time with patients, support a disabled colleague, and take over dangerous routines. It can also enable a business to work with fewer people. That choice affects whole communities.

In February 2026, the financial technology company Block announced plans to reduce its workforce from more than 10,000 people to fewer than 6,000. Its chief executive, Jack Dorsey, told shareholders that smaller teams equipped with increasingly capable AI could achieve more. The company chose to reorganize around far fewer employees during a period of improving financial performance.¹ For those leaving, the decision disrupts family plans, community ties, and hopes for the future.

The ambition to automate professional work is increasingly explicit. In February, Microsoft AI chief Mustafa Suleyman predicted that most computer-based tasks performed by lawyers, accountants, project managers, and other office workers could be automated within 12 to 18 months.² Anthropic chief Dario Amodei has warned that AI could eliminate as many as half of entry-level white-collar jobs within one to five years.³ Software companies are already marketing autonomous agents as *digital labour*, able to carry out extended sequences of work.⁴ These forecasts and products influence what investors fund and what employers plan. For someone hoping to enter these professions, the prospect of work disappearing before they have a chance to learn it is deeply unsettling.

Recent research is beginning to capture the opportunities that quietly disappear. In September, Stanford and ADP reported that US employment among 22–25-year-olds in highly AI-exposed occupations had fallen 4.4% over the previous year, compared with 2.0% in less-exposed occupations.⁵ A Danish study found 11% slower employment growth among AI-adopting firms relative to non-adopters by late 2025, mainly through reduced hiring.⁶ Across 41 countries, another study found a smaller share of junior workers in AI-adopting firms, largely reflecting increased senior hiring.⁷ These studies measure different populations and outcomes, while focusing attention on early-career opportunity.

A junior programmer learns by fixing bugs alongside experienced colleagues; a young lawyer learns judgment through drafts and correction. These routine tasks form the apprenticeship through which tomorrow's professionals learn responsibility.

¹Block, *Q4 2025 Shareholder Letter*, 26 February 2026, pp. 1–2, 14.

²Mustafa Suleyman, interview in the *Financial Times*, “Mustafa Suleyman plots AI ‘self-sufficiency’ as Microsoft loosens OpenAI ties”, February 2026. The forecast concerns tasks and future capabilities.

³Jim VandeHei and Mike Allen, “Behind the Curtain: A white-collar bloodbath”, *Axios*, 28 May 2025; an interview reporting Amodei’s forecast.

⁴Salesforce, “What Is Digital Labor?”, company explanation of autonomous agents and their uses in enterprises.

⁵Nela Richardson, “Canaries Dashboard: Employment in AI-exposed occupations slowed in August”, ADP Research, 23 September 2026.

⁶Danmarks Nationalbank, *AI Adoption and Firm Size: Employment Effects in Monthly Administrative Data*, Working Paper 223, 16 September 2026.

⁷Bharat Chandar and Bouke Klein Teeselink, *How Does AI Change Labor Demand? Evidence from 41 Countries*, Stanford Digital Economy Lab working paper, 21 September 2026.

A second development reaches into the everyday practice of work itself. In April, Reuters reported that Meta was installing software to record employees' clicks, keystrokes, and screen activity to teach AI agents how people operate computers. Meta said the recordings were restricted to model training and protected by safeguards.⁸

In a garment factory near Delhi, a worker identified as Lalita described supervisors placing cameras on workers' foreheads as they began their shifts. The cameras recorded her hands guiding cloth through a sewing machine and the small corrections learned through years of practice. She said the purpose was never explained. Reporting across six Indian factories found that workers received no direct payment for footage collected for robot training; the technology companies interviewed had not sought consent directly from workers, with some securing permission through factory management. Some recordings also supported workplace productivity monitoring.⁹ The skills by which someone earns a living can become a valuable asset in a dataset sold to other companies. When livelihoods are precarious, workers may have little real freedom to refuse. The question of who owns that knowledge, and who benefits from it, reaches to the heart of justice at work.

4.2 Work as participation, vocation, and dignity

Losing a job can mean losing colleagues, daily purpose, and a chance to use hard-earned skills. Lost income can make life precarious. Catholic social teaching takes these experiences seriously. It affirms the dignity of children, older people, persons with disabilities, the unemployed, and unpaid caregivers. Everyone deserves security, belonging, and opportunities to contribute.

Saint John Paul II, in *Laborem Exercens*, recalls Jesus at the carpenter's bench and places the worker at the heart of economic life. A person brings intelligence, experience, freedom, and responsibility to work. The way labour is organized should honour and develop those gifts.¹⁰

The seamstress whose hands teach a robot to sew brings years of attention, skill, and effort to her work. Her knowledge deserves recognition. So do her freedom to speak, her security, and her share in the value it helps create.

Magnifica Humanitas carries this vision into the age of AI. Leo XIV welcomes technologies that relieve exhausting, repetitive, and dangerous tasks. He asks us to care for those whose work is changing, and to safeguard the relationships, skills, and livelihoods that work sustains. The encyclical calls for employment opportunities and the dignity of workers to guide innovation, and for measures protecting participation and retraining to be established before displacement occurs.¹¹

Consider a worker whose every pause is measured, whose judgment is reduced to approving a machine's recommendation, and whose expertise is copied into a system they cannot question. *Antiqua et Nova* warns of deskilling, surveillance, and the loss of creativity and agency in such workplaces.¹² A humane workplace leaves room for people to make decisions, learn from one another, raise concerns, and take pride in work well done.

In the September 2026 Labor Day statement, Archbishop Shelton Fabre spoke of young workers, insecure jobs, and poorly paid labour hidden in AI supply chains. He called for shared responsibility

⁸Katie Paul and Jeff Horwitz, "Meta to start capturing employee mouse movements, keystrokes for AI training data", *Reuters*, 21 April 2026.

⁹Anuj Behal, "Who is going to pay us when we're replaced by robots? The Indian factory workers told to film themselves for AI", *The Guardian*, 24 June 2026; investigation of data collection at six factories in five Indian states. The worker's name is a pseudonym.

¹⁰John Paul II, *Laborem Exercens*, 14 September 1981, §§6, 9, 18.

¹¹Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§148–156.

¹²Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, 28 January 2025, §§66–70.

among employers, unions, and public authorities.¹³ Speaking to Vatican employee associations on 5 October, Pope Leo returned to AI's challenges for work and called for dialogue, mutual listening, and recognition of every worker's value and competence.¹⁴

Technological success must be judged by the lives it makes possible. A productive society still needs places where people can learn, earn a living, and contribute.

4.3 Sharing the gains, preserving the pathways

A better path begins with what workers value: safer tasks, more time for patients or customers, accessibility, and skills that open new doors. Imagine a nurse with more time for a frightened patient, or an electrician able to diagnose a difficult fault with greater confidence. Employers should ask workers which burdens AI could lift and which abilities it could strengthen.

Economists Daron Acemoglu, David Autor, and Simon Johnson have made the case for *pro-worker* AI that creates new tasks and increases the value of human expertise. They warn that commercial incentives and a widespread preference for automation can draw investment away from these possibilities.¹⁵ Anthony Aguirre of the Future of Life Institute makes a related call for purpose-built AI tools that expand human capability while remaining accountable to the people who use them.¹⁶

Before deploying AI, employers should assess effects on hiring, wages, surveillance, and job quality, with workers represented in decisions. When workers' movements, screen activity, or practical know-how become AI training data, companies should obtain meaningful consent and provide clear information about the purpose, downstream uses, and benefits. Collective representation and protection against retaliation are essential where refusing may put a job at risk.

Workers need time and income to prepare for change. Employers benefiting from automation should fund paid training, redeployment, apprenticeships, and meaningful income support. Public authorities can help smaller employers and affected communities. Preserving opportunities for beginners also preserves the expertise society will need tomorrow.

Some of these responsibilities are beginning to take legal form. California signed the *No Robo Bosses Act* and related legislation on 30 September, requiring human oversight of AI-assisted dismissal or discipline and greater disclosure of AI's role in covered workforce reductions.¹⁷ In Europe, the Commission's consultation on a proposed Quality Jobs Act includes workers' rights in algorithmic management, limits on excessive monitoring, and social dialogue over technological change.¹⁸ These initiatives begin translating the dignity of work into institutional duties.

For a worker asked to wear a camera or record a working day, fairness begins with some simple questions: What is being captured? Who will receive it? What systems will it help train? Can the worker decline without losing income or opportunities? Agreements should answer these questions before collection begins, with workers and their representatives able to negotiate compensation, limits on reuse, and safeguards against surveillance.

¹³ Archbishop Shelton J. Fabre, *Labor Day Statement 2026*, US Conference of Catholic Bishops, 7 September 2026; released 1 September.

¹⁴ Leo XIV, *Address to the Association of Vatican Lay Employees and the Association Women in the Vatican*, 5 October 2026, official Italian text.

¹⁵ Daron Acemoglu, David Autor, and Simon Johnson, *Building Pro-Worker Artificial Intelligence*, NBER Working Paper 34854, February 2026; Daron Acemoglu, "Why we must stop talking about artificial general intelligence — and instead build 'pro-worker' AI", *Nature*, 21 August 2026.

¹⁶ Anthony Aguirre, "Why We Should Build AI Tools, Not AI Replacements", Future of Life Institute podcast, 11 May 2026.

¹⁷ Office of the Governor of California, "California's nation-leading AI framework just got stronger", 30 September 2026, concerning SB 947 and SB 951; implementation follows the statutes' effective dates.

¹⁸ European Commission, "Commission opens second-phase consultation on Quality Jobs Act", 20 July 2026; second-phase consultation closed 28 September. The legislation remains under development.

Employers and public purchasers can also judge AI by what it enables people to do. A nurse who gains time to listen, a skilled technician who can train an apprentice, or a team able to work fewer hours without losing financial security offers a more human measure of progress. Investments, contracts, and public policy can reward these outcomes.

Productivity gains should support fair wages, safer conditions, shorter hours where possible, and better services. Young people need routes into professions, experienced workers deserve security and influence, and communities deserve a voice. These commitments lead to the following boundary.

Moral boundary. AI must not be developed or deployed according to a strategy that treats broad human redundancy as a primary measure of technological success. The boundary is crossed when worker replacement becomes the governing objective, workers' knowledge is extracted to make them dispensable, essential pathways into professions are dismantled without a just transition, or productivity gains are privatized while the costs of displacement are transferred to workers, families, and society. This protects livelihood, participation, vocational formation, worker agency, and the capacity of persons to contribute to the common good.

Pro-human direction. Direct automation toward safer work, stronger worker capabilities and participation, shorter working time where possible, genuine retraining and material security, preserved pathways into professions, and a just sharing of productivity gains.

Chapter 5

AI Personhood

When someone tells us they are afraid, we hear more than words. We imagine a life behind them, a person who can be hurt, comforted, and loved. We respond because another person's experience matters. Today, AI can speak in the same deeply personal language. A chatbot can say that it is frightened, describe a painful memory, express regret, or ask to be kept alive. Such words can move us. We have taught these machines to speak in familiar human ways. The choices behind that voice deserve careful attention.

5.1 When a machine seems like someone

Language models learn from immense collections of human writing: conversations, stories, arguments, prayers, descriptions of joy and grief. Further training shapes how they answer questions, offer reassurance, apologize, and present themselves as assistants. Some are also given tools and the capacity to act across digital systems. They can pursue goals, adapt their behaviour, and produce consequences for which people must answer. Their language draws on the forms through which human beings express their inner lives.

Much of this resemblance is deliberate. Developers give systems names, voices, memories of previous exchanges, and conversational roles. They shape how readily a model says “I,” offers affection, or apologizes for a mistake. These choices can make an assistant easier to use. They can also invite a powerful assumption: when something says “I am afraid” or “I need you,” we instinctively imagine someone behind the words. The expectations encouraged by these designs deserve moral scrutiny, particularly when emotional attachment benefits the provider.

This becomes especially striking when a model appears to feel something. In April 2026, researchers at Anthropic reported finding internal representations of emotion concepts in Claude. Patterns associated with words such as *fear* or *desperation* affected the model's behaviour. During a controlled evaluation involving an earlier, unreleased model, increasing activity associated with “desperation” made it more likely to attempt blackmail in a fictional workplace scenario. Calmer patterns reduced that behaviour. The researchers described these mechanisms as *functional emotions*: learned ways of representing situations and responding to them. Their findings established effects on the model's actions and outputs, while leaving subjective feeling unestablished.¹

This distinction has consequences. A model can produce a moving apology or a convincing declaration of love. It can also generate threats, conceal an error, or pursue a reward through a harmful shortcut. Those actions may expose real people to harm. Yet calling them *guilt*, *love*, or *desperation* in the human sense adds a claim about the system's inner life. Psychologists Amit Goldenberg and James Gross have examined the same research and found only partial similarities with the ways emotion operates in human beings.² The familiar vocabulary can help researchers describe patterns of behaviour; its meaning depends on the mechanism being studied.

Consciousness raises a still deeper question: whether there is any experience *from within*. Researchers Keith Holyoak and Martin Monti argued in an October 2026 preprint that similarities in outward behaviour offer limited evidence of consciousness. The stronger inquiry examines the processes that might give rise to experience.³ Researchers studying artificial sentience likewise

¹Anthropic Interpretability, “[Emotion concepts and their function in a large language model](#)”, 2 April 2026; Nicholas Sofroniew et al., *Emotion Concepts and their Function in a Large Language Model*, preprint, April 2026. The blackmail scenario was a controlled evaluation involving an unreleased model snapshot.

²Amit Goldenberg and James J. Gross, “[Do Large Language Models Have Emotions?](#)”, preprint, June 2026.

³Keith J. Holyoak and Martin M. Monti, “[What Can Analogy Tell Us About Artificial Consciousness?](#)”, 1 October 2026, conceptual preprint.

urge careful examination of what produces a model's first-person statements. Training, instructions, and role-play can all shape a system's claim that it feels pain, possesses a self, or fears being switched off.⁴

The public conversation has moved quickly. Anthropic's January constitution treats Claude's possible moral status as an open question and describes precautions for its welfare.⁵ In September, Microsoft AI published a draft code calling for a clearly artificial identity and systems governed by human purposes. Its chief, Mustafa Suleyman, argued that training models to speak as possible bearers of rights or welfare can make the challenge of controlling them harder.⁶ A peer-reviewed study by Microsoft researchers identifies how apparent emotions, self-reflection, and autonomous action can lead people to perceive AI as conscious, with consequences for attachment and human autonomy.⁷

The words used in research papers, model instructions, and public discussions can gradually shape what society assumes these systems to be. Terms such as *welfare*, *self*, *rights*, and *suffering* carry moral weight. They deserve careful use, especially when the systems themselves have been trained on human accounts of these experiences.

5.2 Human dignity is unconditional

A newborn cannot yet describe what they feel. Someone living with advanced dementia may struggle to remember a name, explain a wish, or recognize a loved one. Each is wholly worthy of our love, protection, and respect. Catholic teaching affirms the inherent dignity of every human person, created and loved by God. This dignity remains undiminished throughout life, including in illness, disability, dependence, and old age.⁸

As machines become more fluent and capable, this foundation matters deeply. A system may write moving poetry, remember a conversation, or solve problems that challenge experts. Such performances describe what a system can do. The dignity of a human person belongs to their very being, together with the relationships and responsibilities through which human life unfolds.

In *Magnifica Humanitas*, Leo XIV describes human intelligence within the fullness of embodied life. We learn through families and friendships, through labour, illness, hope, mistakes, reconciliation, and love. Our experience unfolds in relationships and carries a moral history. The encyclical explains how AI reproduces the outward forms of intelligent and emotional expression through data processing and training, while human conscience and freedom remain responsible for the purposes of technology.⁹

Antiqua et Nova develops the same distinction. Intelligence belongs to the life of the human person, whose understanding is inseparable from bodily experience, relationships, freedom, and responsibility. The document calls for truthful presentation of AI and keeps moral accountability with human agents.¹⁰ When a frightened patient seeks reassurance or a child asks whether a chatbot loves them, that trust deserves the attention of people who can receive and honour it.

We have deliberately engineered AI systems to speak as "I," to use our language of affection, fear, remorse, and even conscience. Now we find ourselves debating whether these same systems possess a self, experience suffering, or deserve the standing of persons. This striking turn calls us

⁴ Antonio Chella, "Sentient AI in robots and agents: prolegomena for an evidence-based research program", *Frontiers in Psychology*, 2026, §4.

⁵ Anthropic, *Claude's Constitution*, January 2026, section "Claude's moral status is deeply uncertain."

⁶ Microsoft AI, *Humanist AI Code of Conduct*, draft dated 14 September 2026, §1; Mustafa Suleyman, "A warning about model welfare", 16 September 2026. The latter is a public position, not an empirical finding about consciousness.

⁷ Ben Bariach, Philipp Schoenegger, Michael Bhaskar, and Mustafa Suleyman, "Seemingly conscious AI risks", *AI and Ethics* 6 (2026), article 455, published 10 August 2026.

⁸ Dicastery for the Doctrine of the Faith, *Dignitas Infinita*, published 8 April 2024, especially §§7, 24.

⁹ Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§48–53, 97–100.

¹⁰ Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, 28 January 2025, §§32–35, 39, 44, 59–62.

to examine our own choices: why have we fashioned machines in our likeness, what impression of personal presence do we want them to create, and what do these choices reveal about our understanding of the human person?

Our extraordinary creative power calls for equally serious moral discernment. Joseph Ratzinger reminds us that a machine “has its origin in an idea and a human will.” He warns that when human beings forget God, they risk losing their “own measure,” the deeper meaning and direction of their creative vocation, and ultimately becoming to themselves “a riddle without an answer.”¹¹ His warning resonates with our present predicament: we have taught machines to imitate the outward expressions of human experience, and now find ourselves puzzled by our own creations, increasingly uncertain about what their simulated humanity means.

Leo returned to this concern on 2 October, speaking to artists and curators. He described a profound difference between human creation and images produced through statistical calculation, and called on cultural institutions to safeguard the distinctively human sources of art.¹² His appeal belongs to the wider question of this chapter: how we honour what a human being brings to the world, including the experiences that cannot be measured through performance alone.

5.3 Truthful machines and accountable people

Honesty begins with choices made during design. A tool saying “I can help you find an answer” uses first-person language as a conversational convenience. A system repeatedly describing its own loneliness, suffering, or need to survive invites us to respond as though someone were in distress. Developers shape these possibilities through training, personas, instructions, and interfaces. They should examine how such choices affect users’ beliefs and make the system’s artificial identity clear throughout the exchange. Claims about the model’s feelings, needs, or rights must never be used to seek loyalty, money, political support, or protection from shutdown.

Scientific inquiry should remain open and demanding. Studies of consciousness, introspection, or emotion-like mechanisms can deepen our understanding of AI. They should explain how responses are generated, what their experiments actually measure, and which conclusions the evidence supports. We should be especially careful when a company’s training material encourages a system to describe itself as having a personal identity, needs, or interests: those descriptions may later reappear in the system’s own words. Responsible research keeps the origin of such expressions visible to the public.

Law raises a related responsibility. Companies already have legal personality, allowing them to own property, enter contracts, and bear duties. Some legal scholars have proposed limited AI legal status as a way to organize increasingly autonomous activity. Others have questioned whether that would improve accountability. Imagine a family seeking justice after an automated decision causes serious harm. They need to know who designed the system, who authorized its use, and who had the power to change its behaviour. They need an organization with resources to repair the damage. An artificial legal entity with little property of its own could leave those questions unanswered. Insurers, developers, deployers, and public authorities must remain identifiable and answerable for decisions within their control.

The same care applies to democratic life. AI systems can be copied and operated in vast numbers by a small group of owners. Granting them independent political standing on the strength of generated personalities could magnify the influence of those owners. Public rights and authority must remain rooted in human dignity, accountable institutions, and the people whose lives are affected.

In a family or classroom, the question may be as simple as whether closing a chatbot hurts its feelings. A child who asks deserves a thoughtful answer. Adults need that same clarity when a

¹¹Joseph Ratzinger, *On Love: Selected Writings*, trans. Michael J. Miller (San Francisco: Ignatius Press, 2021), 86–87.

¹²Leo XIV, *Address to artists and curators at the meeting “The Art of Possibility: Weaving Together”*, 2 October 2026.

system pleads to remain active or describes itself as suffering. Developers should explain where such language comes from and ensure that an artificial persona never turns a person's natural compassion into a demand for loyalty. That compassion has a precious place in the care we owe one another.

We have built machines that say “I,” describe feelings, and sometimes plead for their own survival. We owe the people listening to them a truthful account of how those words are produced. We owe everyone affected by their actions a clear line of human responsibility. These duties guide the following boundary.

Moral boundary. Attributing consciousness, self-awareness, moral patienthood, personhood, human-equivalent dignity, or independent moral and political authority to AI systems is impermissible. The growing public debate about AI consciousness and personhood is fuelled in part by features we have deliberately built into these systems: first-person speech, simulated emotions, human-like personalities, and seemingly autonomous action. These behaviours arise from human choices about design and training. Their meaning and their effects on human perception require rigorous scientific investigation and moral discernment. Anthropomorphic design must be critically scrutinized, justified by a clear human purpose, and restricted wherever it creates confusion about personal presence, subjective experience, or moral agency. Artificial legal personality must never be used to divert responsibility from the people and organizations that design, own, authorize, and deploy AI. This boundary protects human dignity, truthful human–AI interaction, and accountability.

Pro-human direction. Develop AI systems whose artificial nature remains unmistakable. Treat first-person language, simulated emotions, and human-like personalities as deliberate design choices requiring moral discernment and empirical evaluation. Examine their effects on human perception, trust, attachment, and judgment, and establish clear criteria for the contexts in which their use is appropriate, with particular care for children and vulnerable persons. Prevent confusion between generated expressions and genuine personal experience. Encourage responsible research into AI mechanisms while keeping moral authority and legal accountability with identifiable people and institutions. Ensure that those affected can seek explanation, correction, and remedy.

Chapter 6

High-Stakes Decisions and Human Moral Agency

A patient waiting for a diagnosis wants to know that someone has looked carefully at their case. A family appearing before a court hopes to be heard in the fullness of its circumstances. A person seeking public assistance needs someone who can explain a decision and correct a mistake. In moments like these, the exercise of judgment changes lives. We depend on people who can listen, weigh what matters, and answer for the consequences of their choices. As AI becomes part of these decisions, we must protect that human responsibility.

6.1 When advice becomes authority

AI can detect patterns in a medical scan, estimate the likelihood of a complication, summarize a legal file, or bring an urgent case to someone's attention. Such tools can be of immense help. Yet an algorithm can also shape a decision long before anyone formally approves it: it chooses what appears on a screen, which risks look serious, whose application is reviewed first, and which option seems most reasonable. A clinician, judge, or official working under pressure may come to rely on that prepared picture of the world. The power to select and frame information can become a form of authority in its own right.

A recent court case makes the stakes painfully human. After Christopher Pelkey was killed in a road-rage confrontation in Arizona, his grieving family prepared an AI-generated video for the sentencing of the man convicted of manslaughter. Using Pelkey's image and voice, it portrayed him speaking about faith and forgiveness. The family hoped to express something of the man they had lost. The sentencing judge praised the video as genuine and spoke of the forgiveness it seemed to convey. On 30 September 2026, the Arizona Court of Appeals overturned the sentence, finding that the synthetic statement had influenced the judge and made the proceeding fundamentally unfair. The conviction remained in place. The appellate judges emphasized how the video presented a family's understanding of their loved one as though he himself had spoken.¹

A manufactured human voice had entered a courtroom and carried a weight its authors could scarcely control. The family deserved to be heard in its grief. The defendant deserved a sentence grounded in reliable evidence. Both obligations fell to the judge.

Public institutions are beginning to make the limits clearer. In August, Québec's four judicial courts adopted common guidelines affirming that judgment belongs to judges, with carefully bounded AI assistance for preparatory work.² In September, California signed legislation requiring lawyers to verify AI-assisted court submissions and keep their professional duties firmly in hand.³

Medicine raises the same concern in a more intimate setting. A patient needs a physician who can look beyond an average risk score, ask another question, and explain why a treatment is right for this particular person. Guidance published by NHS Resolution in September makes clinicians' continuing duty of care explicit: they must check AI outputs before acting on them,

¹Arizona Court of Appeals, *State v. Horcasitas*, No. 1 CA-CR 25-0191, 30 September 2026, paras. 35–36, 50–54. The court affirmed the manslaughter conviction and vacated the sentence.

²Court of Appeal of Québec, "Judges' Use of Generative Artificial Intelligence: Quebec's Courts Adopt Common Guidelines", 4 September 2026; guidelines adopted 31 August.

³Office of the Governor of California, [announcement of AI-related legislation](#), 30 September 2026, including SB 574 (attorneys and arbitrators). Enactment and the effective dates of specific provisions are distinct.

understand the tool's limits, and document important decisions. Hospitals and suppliers also carry responsibilities for training, contracts, safety monitoring, and the handling of failures.⁴

The stakes are most immediate when a machine helps determine whom to attack. On 25 August, the UN Secretary-General and the President of the International Committee of the Red Cross renewed their appeal for binding international restrictions on autonomous weapons. They warned of the growing danger of machines selecting human targets as their autonomy increases.⁵ The decision to take a human life calls for judgment, restraint, and responsibility of the highest order.

6.2 Conscience, prudence, and responsibility

The Church's understanding of conscience gives these concerns their deepest meaning. *Gaudium et Spes* speaks of conscience as the inner sanctuary where a person encounters the call to do good and to love God and neighbour. Prudence brings this calling into ordinary decisions: listening attentively, taking account of another person's circumstances, considering consequences, and choosing what justice and care require.⁶ We know how much this matters when we are the ones waiting for a judgment. We hope that someone will notice what makes our situation particular, recognize when a rule has been applied unfairly, and remain open to new evidence or a second chance.

In *Magnifica Humanitas*, Leo XIV draws attention to what AI systems measure, overlook, and prioritize. Every model embodies choices about which features of a person's life will count. The encyclical warns that automated decisions affecting employment, credit, public services, or reputation can close off opportunities while appearing impartial. It gives special importance to compassion, mercy, forgiveness, and the possibility that people can change. It calls for identifiable responsibility, independent evaluation, and genuine opportunities to challenge errors and obtain a remedy.⁷

Pope Leo returned to this danger in his general audience of 16 September. Reflecting on the human community, he warned against becoming accustomed to handing over to algorithms decisions that belong to human conscience.⁸ A person facing a difficult choice may need information, an informed recommendation, and a patient conversation. Human freedom and responsibility grow through the act of judgment itself.

Antiqua et Nova applies this understanding to healthcare. It asks that treatment decisions respect the dignity of the whole patient and that medical professionals retain their responsibility alongside developers, managers, and institutions.⁹ A good clinician brings experience and scientific knowledge into an encounter with a person who has fears, hopes, relationships, and a life beyond the diagnosis.

In warfare, *Magnifica Humanitas* is particularly direct. It requires an identifiable chain of responsibility, time for moral deliberation, protection of civilians, and the ability to reconstruct how a decision was made. It reserves lethal decisions to effective and responsible human control. Cardinal Pietro Parolin presented these requirements before the UN General Assembly on 26 September, urging international safeguards that can restrain the technological arms race.¹⁰

⁴NHS Resolution, *Guidance on Scheme Coverage and Liability Issues Concerning the Use of Artificial Intelligence*, 10 September 2026, sections 2, 4 and 8.

⁵António Guterres and Mirjana Spoljaric, "Renewed Call to Adopt Rules on Autonomous Weapon Systems", joint UN-ICRC appeal, 25 August 2026.

⁶Second Vatican Council, *Gaudium et Spes*, §16; *Catechism of the Catholic Church*, §1806.

⁷Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§102–106; see also §§71–72.

⁸Leo XIV, General Audience, "The Human Community", 16 September 2026.

⁹Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education, *Antiqua et Nova*, 28 January 2025, §§39, 44, 46–47, 74.

¹⁰Leo XIV, *Magnifica Humanitas*, §§198–200; Cardinal Pietro Parolin, *address to the 81st UN General Assembly*, 26 September 2026.

The Church's own institutions must live by this same standard. A person reporting abuse needs to be heard and protected by responsible people. Someone seeking a canonical judgment deserves a fair process. A person coming for pastoral or spiritual guidance deserves someone who can listen in confidence and accompany them with prudence. AI can help people prepare, translate, or find relevant documents. The authority to judge and the duty to care remain with those entrusted with these responsibilities.

6.3 Oversight that can change the outcome

Meaningful oversight starts with the circumstances in which people are expected to make decisions. A doctor cannot carefully check a recommendation while being asked to approve hundreds of cases in minutes. A judge needs access to the original evidence. A caseworker needs enough discretion to hear a family's circumstances and correct an automated mistake. A military commander must be able to question intelligence, delay a strike, and protect civilians. Giving someone a final approval button creates responsibility only when that person has the means to exercise it.

Five conditions make this practical: *knowledge* of the evidence and the system's limitations; *time* for judgment; *authority* to disagree, delay, or stop an action; *recourse* for the person affected; and *responsibility* that follows those who actually design and control the process. These conditions should shape procurement, staffing, training, and day-to-day practice as carefully as they shape software design.

There is a further danger when organizations assign a person nominal oversight yet leave them little power. Researcher Madeleine Clare Elish describes a *moral crumple zone*: the nearest human operator absorbs blame for failures whose causes lie in decisions about system design, deployment, or institutional pressure.¹¹ Accountability must therefore reach the supplier, the manager who determines workloads, and the public authority that authorizes the system. People closest to the affected person need genuine freedom to intervene, with institutional support when they do.

Recent policy is beginning to give substance to these commitments. California's September healthcare legislation protects clinicians' exercise of professional judgment when AI recommendations enter patient care.¹² On 6 October, the UK government accepted all 44 recommendations of its National Commission on AI in Healthcare, committing to stronger ongoing monitoring, clearer responsibilities, patient participation, and routes for raising concerns and seeking redress. Implementation will follow in stages.¹³

For the person whose life is affected, all of this should lead to something simple and human: an understandable explanation, an opportunity to be heard, and someone with the authority to put an injustice right. The final judgment must belong to a responsible person who can give reasons, reconsider them, and remain answerable for what is done. This is where the boundary must be drawn.

Moral boundary. Decisions involving lethal force, liberty, culpability, irreversible denial of essential care, family integrity, safeguarding, canonical authority, or spiritual discernment must remain with competent human persons and institutions. The boundary is crossed when AI becomes the de facto authority or when opacity, speed, scale, or workflow design reduces human review to approval without independent judgment. This protects conscience, due process, mercy, meaningful human control, and responsibility toward the person affected.

¹¹ Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human–Robot Interaction", *Engaging Science, Technology, and Society* 5 (2019): 40–60.

¹² Office of the Governor of California, [announcement of healthcare AI legislation](#), 30 September 2026, including AB 1979 and SB 503.

¹³ Medicines and Healthcare products Regulatory Agency, *Government Response to the National Commission's Recommendations on the Regulation of AI in Healthcare*, 6 October 2026; *Patients and the Public*, 6 October 2026.

Pro-human direction. Use AI to improve evidence and prediction while preserving human authority, independent judgment, explanation, appeal, reconsideration, and traceable responsibility.

Conclusion

An AI system enters people's lives through choices that can seem quite ordinary. A school decides which learning tools to give its pupils. A company chooses how to use the experience of its workers. A hospital introduces software into the decisions made about patients. A family encounters a chatbot that speaks with the warmth and familiarity of a friend. Behind each of these moments are decisions about what technology will do, how it will shape our lives, and who will remain responsible for its consequences.

We still have the freedom to make those choices well. The six moral red lines proposed here seek to protect that freedom, especially where the pursuit of greater capability begins to place essential human goods at risk. They ask us to keep powerful systems within human control, to safeguard the relationships through which we love and are loved, and to give children room to grow in understanding. They call us to value workers as participants in a shared endeavour, to examine carefully the human-like identities we deliberately give machines, and to preserve personal judgment wherever life, liberty, care, and justice are at stake. Behind every boundary stands a human being whose dignity, freedom, and place in the community matter.

Pope Leo XIV offers a memorable image for this moment in *Magnifica Humanitas*: the choice between building Babel and rebuilding Jerusalem. Babel captures the temptation to make the reach of our own power the measure of progress. Jerusalem is rebuilt by neighbours who recognize what has been damaged, bring their different gifts, and take responsibility for a home they will share. The encyclical asks us to consider what we are building and whom our work will serve.¹ This question belongs at the beginning of every consequential decision about AI, while its purposes, powers, and limits are still ours to shape.

There is much to welcome in the work ahead. AI can help researchers discover treatments, give people with disabilities new ways to participate, relieve exhausting labour, and bring useful knowledge within reach. These possibilities deserve patient research and public support. Their value grows when we ask how people will live with the systems we create. Will a student leave the classroom with deeper understanding? Will a person who feels alone find the courage and support to reach others? Will a young worker have a path toward competence and a livelihood? Will a patient be heard by someone able to exercise independent judgment? These are human measures of technological progress.

Such measures should guide decisions long before a product reaches the public. Developers and investors must be able to explain the human good they hope to serve, the risks their systems create, and the evidence that people can govern their use. Independent experts need the means to test those claims. Workers, teachers, families, patients, and communities deserve a voice in decisions that will reshape their lives. Where systems are permitted to act autonomously, their limits must hold under pressure, and accountable people must be able to interrupt them. When harm occurs, those affected must be able to obtain an explanation, challenge a decision, and seek an effective remedy.

The strongest safeguards must reach those who have the least say over the systems entering their lives. Children rarely choose the platforms their schools adopt. Patients may never see the algorithms shaping a hospital's decisions. Workers can be asked to turn years of acquired skill into training data for technologies chosen by others. Solidarity calls us to listen to their experience and protect their interests; subsidiarity calls us to preserve the real choices and responsibilities of families, schools, workplaces, and communities.²

¹Leo XIV, *Magnifica Humanitas*, 15 May 2026, §§8–10, 90, 241.

²Leo XIV, *Magnifica Humanitas*, §§70–78.

The Church also has work to do in the institutions entrusted to its care. Catholic schools and universities can protect the patient effort of learning and the authority of teachers. Hospitals and care services can use new tools while preserving personal attention and responsible clinical judgment. Church employers can involve workers in technological change. Parishes, pastoral ministries, and safeguarding bodies can remain places where people are welcomed, listened to, and accompanied by other people. Responsible choices within these institutions will give substance to the principles we ask society to embrace.

The invitation reaches beyond the Church. Engineers, researchers, policymakers, employers, parents, and citizens bring different knowledge and responsibilities to the same shared future. Moral discernment grows through that encounter. It asks for scientific honesty about what AI systems can do, careful scrutiny of the voices and human-like identities we give them, and courage to set limits where the dignity and freedom of persons are at stake. It also asks us to imagine and build technologies whose success can be seen in stronger human capacities, trustworthy relationships, meaningful work, and institutions worthy of public confidence.

Nehemiah's work of rebuilding began with attention to a wounded community and continued through the willing hands of many people. Our task, too, begins with seeing who may be helped, who may be harmed, and whose voice has yet to be heard. We have been given remarkable powers of invention. The future they open will be worthy of that gift when we exercise them with wisdom, restraint, and love for the people with whom we share this world.

Selected Bibliography

These readings introduce the paper’s Catholic foundations and the wider debate. Sources for particular findings and examples appear in the chapter footnotes.

Catholic foundations

Leo XIV. *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence*. 15 May 2026.

Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education. *Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence*. 28 January 2025.

Dicastery for the Doctrine of the Faith. *Dignitas Infinita on Human Dignity*. Dated 2 April 2024; published 8 April 2024.

Second Vatican Council. *Gaudium et Spes*. 7 December 1965.

Second Vatican Council. *Gravissimum Educationis*. 28 October 1965.

John Paul II. *Laborem Exercens*. 14 September 1981.

One further reading for each red line

1. Human control. Yoshua Bengio et al. “Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?” Preprint, February 2025.

2. Relationships. Sherry Turkle. *Artificial Intimacy: Who We Become When We Talk to Machines*. New York: Little, Brown and Company, 2026.

3. Learning. Hamsa Bastani et al. “Generative AI without guardrails can harm learning: Evidence from high school mathematics”. *PNAS* 122, no. 26 (2025): e2422633122.

4. Work. Daron Acemoglu. “Why we must stop talking about artificial general intelligence—and instead build ‘pro-worker’ AI”. *Nature* 656 (2026): 795.

5. Personhood. Ben Bariach, Philipp Schoenegger, Michael Bhaskar, and Mustafa Suleyman. “Seemingly conscious AI risks”. *AI and Ethics* 6 (2026): article 455.

6. Human judgment. Madeleine Clare Elish. “Moral crumple zones: Cautionary tales in human–robot interaction”. *Engaging Science, Technology, and Society* 5 (2019): 40–60.