VON HÜGEL INSTITUTE
critical catholic inquiry
ANALYSING • IMAGINING • TRANSFORMING

St Edmund's College
University of Cambridge

Fragility and Flourishing: AI and Human Enhancement or Diminishment

Maria Burke
Gianmarco Contino
Gill Goulding, CJ
and Jonathan Warner

VON HÜGEL INSTITUTE WORKING PAPERS 1/2026
ISSN 2055-3730

Von Hügel Institute Working Paper 1-2026 ©
ISSN 2055-3730
[1 May 2026]

Fragility and Flourishing: AI and Human Enhancement or Diminishment

Maria Burke
Birkbeck, University of London, and St Edmund's College, University of Cambridge

Gianmarco Contino
University of Birmingham, and St Edmund's College, University of Cambridge

Gill Goulding, CJ
Regis College, University of Toronto, and St Edmund's College, University of Cambridge

Jonathan Warner
LCC International University, and Von Hügel Institute, University of Cambridge

Abstract

Artificial Intelligence (AI) is rapidly reshaping professional and personal lives. While AI introduces significant improvements and opportunities for technological advancement, its expansion has uncovered challenges that disproportionately affect society's most vulnerable or marginalised groups, and risk exacerbating existing forms of discrimination and inequalities. Drawing on insights from ethical, theological, economic, and healthcare perspectives, this paper offers a thought-provoking overview of the impact of AI, highlighting both the benefits of technological progress alongside the potential threats to human dignity and intrinsic value. AI-driven digital change is increasing public anxiety around data control, widening economic inequalities, and altering the relationship between work and productivity. The integration of AI into healthcare is reshaping clinical decision-making processes, transforming the delicate balance between medical authority and patient autonomy. Furthermore, AI is renewing the dialogue between faith and reason and raising important questions about free will and adequate safeguarding. The authors argue that responsible AI development and governance are essential to preserve and protect human flourishing, equality, and social justice, and call for an interdisciplinary effort in evaluating and assessing AI's growing capabilities in an increasingly divided and polarised global context.

Keywords: Artificial Intelligence (AI); digital transformation; healthcare; human dignity; human fragility; equality.

1. Introduction

This paper arises from an interdisciplinary research initiative of the Von Hügel Institute. Bringing together different disciplinary expertise, participants have met and presented their work on several occasions over the last few years both at St Edmund's College, University of Cambridge, and online. It seemed an opportune time, now, to bring together our joint efforts in the following paper.

Artificial Intelligence (AI) presents an opportunity to significantly enhance many aspects of our individual lives. Despite its vast potential, AI's influence may not be uniformly beneficial, with marginalized groups—including minorities, people with disabilities, and the frail—facing disproportionate challenges. It is apparent that there are still significant algorithmic inequalities in the world which need to be attended to. Billions are still without broadband, access to the internet, without mobile phones, and access to a digital signal. It is important to note also that these inequalities also contain a significant gap in access to resources between men and women, young and old. The standard frameworks for AI algorithms, policies, and metrics frequently overlook the diversity of society's individuals and values, potentially exacerbating pre-existing biases or introducing new forms of discrimination against the most vulnerable or underrepresented groups.[1]

The advent of AI is reshaping disciplines and sparking a surge in interdisciplinary research. The digital transformation of society, however, brings with it conflicts between technological advancement and the preservation of human values. It is critical to address these issues, developing methods to analyse and resolve such conflicts to maintain alignment between AI-driven progress and our core human values.[2] Explicit values of the developers and deploying entities further shape algorithmic outcomes. An organization prioritising profit may design algorithms that reflect such values, potentially at the expense of equity and social justice.[3]

This paper aims to bring together perspectives from an array of interconnected disciplines: ethics, economics, medicine and theology, all concerned with the safeguarding of human dignity, sanctity of life, and fragility. Professor Maria Burke brings AI into conjunction with

[1] See Joy Buolamwini, and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification", *Proceedings of Machine Learning Research*, 81 (2018). Values embedded in machine learning algorithms can manifest in several forms. Bias in training data can lead algorithms to perpetuate those biases in their output. For instance, facial recognition technologies trained predominantly on lighter-skinned individuals may underperform with darker skin tones (I. D. Raji, and J. Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products", *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, (2019). Additionally, the design of algorithms—including variable selection and weighting—can subtly encode the biases of their creators. An algorithm focusing on income and employment status for credit assessments may inadvertently marginalize the economically disadvantaged (C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016). Furthermore, the goals set by an algorithm's objective function, such as error minimization or accuracy maximization, can reflect the designers' priorities, potentially leading to outcomes that emphasize efficiency over fairness (Solon Barocas, and Andrew D. Selbst, "Big Data's Disparate Impact", *California Law Review,* 2016).

[2] Luciano Floridi, and Josh Cowls, "A Unified Framework of Five Principles for AI in Society", *Harvard Data Science Review,* 1(1), (2019).

[3] Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale: Yale University Press, 2021).

the frailty of the human being and considers the very contemporary question of the anxiety surrounding the engagement with AI by examining specific digital concerns. Professor Gill Goulding CJ explores theological and ethical insights through the use of AI as a positive tool, the need for reflection on AI developments and the overriding importance of guarding human dignity. Professor Jonathan Warner brings his economic expertise to the fore in a consideration of the possible reduction of economic disparities through an AI driven economy. Here he considers a possible scenario of less work and more leisure time and what regulation might be necessary in such an AI driven economy. Professor Gianmarco Contino contributes the medical and translational perspective, grounding the discussion in the concrete realities of clinical practice and the fragilities of patients whose data and bodies increasingly interface with algorithmic systems.

2. The Power of AI and the Frailty of the Human Being (Maria Burke)

The first digital concern can be identified as the sense of anxiety concerning a loss of control over personal data. The tension of this concern is that the short-term benefit of AI will provide increasingly personalised products and services, yet the long-term cost to us is that this level of personalisation may foster the process of social fragmentation. Anxiety is, of course, incredibly difficult to measure. Wang and Wang were among the first to devise a solution to this problem with the creation of an "AI Anxiety Scale".[4] Their study collected data from over three hundred respondents in order to develop a standardised tool to "measure the phenomenon of AI anxiety". Annual indexes which measure public sentiment have become more significant and available in the last few years. Examples include AI Anxiety Indexes, AI Readiness Indexes and AI Trust Indexes. All these provide interesting—if subjective—overviews of how the public view the adoption of Artificial Intelligence in society. Kim, Soh and Kadkol analysed a whole series of psychological factors and interventions regarding terminology around "AI Anxiety".[5] Their work centres on "fear of replacement by AI" including privacy concerns, misinformation and AI biases. Multi-faceted solutions are suggested regarding revising educational, technological, regulatory and ethical guidelines. Yurt and Kasarci investigated "AI use motivation"[6] by collecting data from over one thousand university students. Whilst this study produced excellent results and useful data, further research needs to build upon this foundational work. Others, such as Frenkenberg and Hochman collected empirical evidence re data concerns relating to motivation, dependency and anxiety concluding that an "emphasis on

[4] Y. Y. Wang and Y. S. Wang, "Development and Validation of an Artificial Intelligence Anxiety Scale: An Initial Application in Predicting Motivated Learning Behavior" *Interactive Learning Environments* 30, no. 4 (2022): 619–634, https://doi.org/10.1080/10494820.2019.1674887

[5] Jeff Kim, Shrinidhi Kadkol, Itay Solomon, Hyelin Yeh, Jun Young Soh, Theresa M. Nguyen, Jeong Yun Choi, Sophie Lee, Adith V. Srivatsa, Georgie R. Nahass, and Olusola A. Ajilore, "AI Anxiety: A Comprehensive Analysis of Psychological Factors and Interventions," *AI and Ethics*, *AI Ethics, 5* (2025): 3993–4009, https://doi.org/10.1007/s43681-025-00686-9.

[6] Eyüp Yurt and İsmail Kaşarcı, "A Questionnaire of Artificial Intelligence Use Motives: A Contribution to Investigating the Connection between AI and Motivation," *International Journal of Technology in Education* 7, no. 2 (March 2024): 308–325, https://doi.org/10.46328/ijte.725.

the importance of adopting a phased strategic approach to AI in organisations"[7] may be the most appropriate way forward.

Our anxiety, then, around how we choose to act in the new, potentially AI-led society, concerns the invasion of our data space regarding potential misuse by others, of what we consider to be *our own* personal data. This issue is now of critical importance due to the pervasive nature not just of the internet, but the increasing use of social media platforms and the mobility and convenience of AI technology, as the possibility of having technology implanted into the human body become increasingly available. While these procedures may offer life-changing solutions for health-related conditions, questions arise when such technologies are used primarily for convenience rather than therapeutic benefit (for example, there is no longer any need to physically carry bank cards, library cards and similar items, as these can all be accessed via smartphones. However, technology is also available that allows microchips to be implanted into the hand). If the decision to adopt such technology is motivated by practicality rather than medical necessity, there is a risk of crossing a line and moving towards a posthuman condition associated with significant ethical and moral dilemmas.

The second underlying digital concern is that technology is taking considered judgement away from the human being. The tension of this concern is that the short-term benefit of AI will increase the speed and possibilities of data manipulation, yet the long-term cost is the potential loss of our sense of purpose as humans, the paradox of the existential crisis.

Removing judgement decisions from ourselves is for many a deep-seated fear and arises from questions as to whether our humanity, or *essence* of being, can (or should) be replicated. Our humanity is special to us; we have judgement, wisdom, and revel in our sense of individual personhood which will always separate us from a technologically advanced machine, no matter how sophisticated the algorithmic programming. However, as AI is fast developing, these questions are becoming increasingly urgent. Thielscher considers that whilst the concept of dignity is central to human being, as we progress towards increasingly intelligent technology we should consider whether it is moral or "right" to also grant some form of dignity to intelligent information and AI computer systems.[8] This proposition raises emotive awareness of what situations humans can tolerate. It seems somehow abhorrent that our humanity—our unique intrinsic value—should somehow be assigned to what is, in essence, a *product* that we have created. However, the study debates the ethics of this issue quite objectively from a discussion of opinions that robots either have no value or that the rights of future robots will become so all pervading in our lives that their *kind of life* must also be protected by law. Other current studies question whether "the allure of novel augmented abilities is worth the potential cost of losing authentic, unfiltered human interaction".[9] This study proposes a framework, the "Human Flourishing Benchmark" designed to evaluate how technology supports or potentially

[7] Adi Frenkenberg and Guy Hochman, "It's Scary to Use It, It's Scary to Refuse It: The Psychological Dimensions of AI Adoption—Anxiety, Motives, and Dependency," *Systems* 13, no. 2 (2025): 82, https://doi.org/10.3390/systems13020082.

[8] Christian Thielscher, "Dignity as a Concept for Computer Ethics," *AI Ethics,* 5 (2025): 4061–4067, https://doi.org/10.1007/s43681-025-00693-w.

[9] Sebastian Zepf and Mark Colley, "Human Authenticity and Flourishing in an AI-Driven World: Edmund's Journey and the Call for Mindfulness," *arXiv*, June 1 (2025), https://doi.org/10.48550/arXiv.2505.13953

hinders fundamental human capabilities. Human development views are also critical for our survival and for our flourishing. Lengfelder, Tapia and Biggeri are concerned particularly with human development and a need for increased emphasis on a "people centred approach to AI".[10]

The third and final underlying digital concern is about our future as digitization results in first a reduction, and then, potentially, a complete lack of human choices as complex technology replaces human free will. The tension of this concern is that the short-term benefit of AI will increase computing sophistication, yet the long-term cost is that we may lose our free will, and our ability to make decisions for ourselves. The development of bio-technologies, robotics, sophisticated algorithms, advances in new materials and the merging of the disciplines has brought significant changes and progress for humanity, such that we live in a world where many diseases and afflictions may be cured by the application of technology, when applied to good purpose.

The hierarchical systems of many organisations can result in power becoming a central determinant of success. Yet, there is an increasing need and emphasis today on mental health, physical well-being and the pursuit for happiness. Effective working relationships in organisations often depend on collegiality, mutual respect, a shared sense of purpose, and camaraderie. These are important elements that highlight the value of interacting with other humans, rather than with a programmed artificial intelligence system. Human communication remains a vital bond that can reassure us of humanity's continuing central role in organisational life.

Scholars continue to be concerned with the study and refinement of Large Language Models (LLM);[11] with preserving our human judgement; with our human safeguarding; with the ethics of AI, with justice, and with our environment.[12] As AI systems begin to act in increasingly human ways, the need to think deeply about what flourishing means both now and in the future is at a critical point.

3. Theological and Ethical Insights in AI Development (Gill Goulding, CJ)

"Human intelligence is an expression of the dignity with which we have been endowed by the Creator who made us in his own image and likeness […] In a particular way, science and technology manifest this fundamentally relational quality of human intelligence; they are brilliant products of its creative potential".[13] This emphasis on intelligence as an expression of the dignity given to human beings in the very act of their creation, firmly establishes the path

[10] C. Lengfelder, H. Tapia, and M. Biggeri, "Navigating AI with a Human Development Compass – Shaping Tomorrow's Capabilities", *Journal of Human Development and Capabilities*, 26 (3), (2025): 439–448, https://doi.org/10.1080/19452829.2025.2520011.

[11] Gabriel Lau, Wei Yan Low, Seow Min Koh, and Andree Hartanto, "Evaluating AI Alignment in Eleven LLMs through Output-Based Analysis and Human Benchmarking," *arXiv (*2025), https://doi.org/10.48550/arXiv.2506.12617.

[12] L. Mura, L., and B. Stehlíková, "The Ethics of Artificial Intelligence: Safeguarding Human Dignity, Social Justice and Environmental Stability in the Age of AI", *Equilibrium*, *Quarterly Journal of Economics and Economic Policy*, 20(2), (2025): 479–507, https://doi.org/10.24136/eq.3743

[13] Pope Francis, "Message of His Holiness Pope Francis for the 57th World Day of Peace" (1 January 2024), accessed 14 May 2026: https://www.vatican.va/content/francesco/en/messages/peace/documents/20231208-messaggio-57giornatamondiale-pace2024.html

of AI within the field of divine human relations and inter-human relations: "The inherent dignity of each human being and the fraternity that binds us together as members of the one human family must undergird the development of new technologies and serve as indisputable criteria for evaluating them before they are employed".[14] In this way AI might serve our best human potential and our highest aspirations. The following remarks are focused on three major points: the use of AI as a tool; the importance of reflection as AI develops and the importance of safeguarding the inviolable dignity of each human person and respecting the cultural and spiritual riches and diversity of the world's peoples.

3.1. AI as a Positive Tool

There are significant examples of best practices of using AI as a tool in a manner that serves the good of vulnerable populations. A prime example concerns the vital commodity of water. The Food and Agricultural Organization of the United Nations (FAO) has a portal that monitors and reports on agriculture water productivity over Africa and the Near East. It provides open access to the water productivity database, and its thousands of underlying map layers allows for direct data queries, time series analysis, area statistics and data download of key variables associated with water and land productivity assessments. The Agricultural Stress index system is an indicator developed for the early monitoring of water stress/drought at global, regional and country level, using satellite technology. Drought affects more people than any other type of natural disaster and therefore the early warning possibilities can save many lives. In a similar manner in September 2020 in Rome the FAO, IBM and Microsoft came together in an event organized by the Pontifical Academy for Life. Discussion focused on the concrete ways through which AI can contribute to achieving the goal of feeding of an estimated global population of nearly 10 billion in 2050, and to do this while safe-guarding natural resources and addressing challenges such as climate change and the impact of pandemics like Covid-19.

Amidst the many and varied benefits of AI it is important to remember that technological innovations are not disembodied and *neutral* but dependent also on the aims and interests of owners and developers. Accordingly, we cannot presume a commitment on the part of those who design algorithms to act ethically and responsibly. There is a need to establish groups who have the responsibility to examine the ethical issues that are ongoingly arising. Given that now AI involves a wide spectrum of different entities, it is important to see that this expansion is matched by an appropriate formation in responsibility for development. The challenges AI poses are not just technical but also anthropological, educational, social and political.

3.2. The Importance of Reflection

Reflection reminds us that the human mind is never able to know the fullness of the extraordinary vastness and complexity of the planet even with the assistance of the most advanced algorithms. Such algorithms do not offer guaranteed predictions of the future, but only statistical approximations. Yet not everything can be either predicted or calculated, in the

[14] Pope Francis, "Message", 2024.

end as Pope Francis states "realities are greater than ideas".[15] Here, it is important to have a sense of human limitation as creatures who are mortal, a reality generally overlooked in our pursuit of ever more advanced algorithms. Pausing to acknowledge such limitation may positively give a necessary space to enable engagement with the more profound ramifications of human dignity and human existence. This is the space for reflection. In the February 2020 meeting already referred to the Pontifical Academy for Life, Microsoft, IBM, FAO and the Italian government signed a document developed to support an ethical approach to AI and promote a sense of responsibility amongst organizations, governments and institutions with the aim to create a future in which digital innovation and technological progress serve human genius and creativity, and not their gradual replacement.

3.3. Guarding Human Dignity

Our AI practices often rely on categorizing individuals. Yet the fundamental respect for human dignity, indicated in the first point above, demands that we refuse to allow the uniqueness of the human person to be identified with a set of data. Algorithms must not be allowed to determine how we understand human rights or to set aside the essential human values of compassion, mercy and forgiveness, or to eliminate the possibility of an individual changing and leaving his or her past behind. In this connection it is important to draw attention to the use of social media especially by the young. The design of the platforms and the algorithms that dictate their performance can play on the worst of our human tendencies leading to online environments that violate the core Christian values of truth and human dignity. Of their very nature digital relationships do not follow the path of the slow cultivation of friendships that mature over time. In 2009 Pope Benedict stated that "if the desire for virtual connectedness becomes obsessive it may in fact function to isolate individuals from real social interaction while also disrupting the patterns of rest, silence and reflection that are necessary for healthy human development".[16]

In 2020 the Rome Call for AI ethics had three key principles which continue to remain relevant. First an ethical assertion that all human beings are born free and equal in human rights and dignity. Therefore, we need to be mindful of the complex reality of our eco-system, our shared home. Secondly that in the field of education which can transform the world, the innovation of AI means undertaking to build a future for and with younger generations. Here, promoting critical thinking is vital to help students and professionals to grasp the social and ethical aspects of the development and use of technology. The third principle involves the area of rights. Here the development of AI in the service of humankind and the planet necessitates regulations that protect people, particularly the most vulnerable and marginalized. It was stated that AI based technology must never by used to exploit people in any way. A set of key characteristics for *Algorethics* were raised in Rome. These were cited as fundamentals of good

[15] Pope Francis, "Apostolic Exhortation", *Evangelii Gaudium* (24 November 2013): 233, accessed 14 May 2026: https://www.vatican.va/content/francesco/en/apost_exhortations/documents/papa-francesco_esortazione-ap_20131124_evangelii-gaudium.html

[16] Pope Benedict XVI, "Message of the Holy Father Benedict XVI for the 43rd World Communication Day", (24 May 2009), accessed on 14 May 2026: https://www.vatican.va/content/benedict-xvi/en/messages/communications/documents/hf_ben-xvi_mes_20090124_43rd-world-communications-day.html

innovations. They included: transparency, inclusion, responsibility, impartiality, reliability, and security and privacy.

Drafting fundamental guidelines for regulatory frameworks for AI also need to take into consideration the deeper issues regarding the meaning of human existence, the protection of fundamental human rights and the pursuit of justice and peace: "this process of ethical and juridical discernment can prove a precious opportunity for shared reflection on the role that technology should play in in our individual and communal lives, and how its use can contribute to the creation of a more equitable and humane world".[17]

Pope Leo has followed the consistent line of his predecessor in being concerned particularly about the effect of AI on children and young people and the possible consequences on their intellectual and neurological development. No generation, Pope Leo acknowledged, has ever had such quick access to the amount of information now available through AI: "but access to data—however extensive, must not be confused with intelligence, which necessarily involves the person's openness to the ultimate questions of life and reflects an orientation toward the True and the Good".[18] Pope Leo stressed that taking into account the well-being of the human person not only materially, but also intellectually and spiritually means safeguarding the inviolable dignity of each human person and respecting the cultural and spiritual riches and diversity of the world's peoples. "Ultimately, the benefits or risks of AI must be evaluated precisely according to this superior ethical criterion," he said warning against societies today experiencing a certain "loss" or "at least an eclipse" of what is human, and saying this "challenges all of us to reflect more deeply on the true nature and uniqueness of our shared human dignity."[19]

In January 2025 the Dicastery for Culture and Education with the Dicastery for the Doctrine of the Faith published a joint document *Antiqua et Nova* which was an extended note on the relationship between AI and human intelligence. It stated: "Human abilities and creativity come from God and, when used rightly, glorify God by reflecting his wisdom and goodness. In light of this, when we ask ourselves what it means to 'be human' we cannot exclude a consideration of our scientific and technological abilities".[20] Within these parameters the document addresses the anthropological and ethical challenges raised by AI, not least the potential role it might have in the growing crisis of truth in the public forum. In addition, it raises fundamental questions regarding ethical responsibility and human safety and prompts reflection on what it means to be human and the role of humanity in the world.

In November 2025 an AI Forum was held in Rome with the aim of fostering a new interdisciplinary community of practice dedicated to supporting the development of AI

[17] Pope Francis, "Message", 2024.
[18] Pope Leo XIV, "Message of Pope Leo XIV to Participants in the Second Annual Conference on Artificial Intelligence, Ethics, And Corporate Governance", (19-20 June 2025), https://www.vatican.va/content/leo-xiv/en/messages/pont-messages/2025/documents/20250617-messaggio-ia.html
[19] Pope Leo XIV, "Message", 2025.
[20] Dicastery for Culture and Education, and Dicastery for the Doctrine of the Faith, "*Antiqua et Nova*: Note on the Relationship between Artificial Intelligence and Human Intelligence" (2025), accessed 14 May 2026: https://www.vatican.va/roman_curia/congregations/cfaith/documents/rc_ddf_doc_20250128_antiqua-et-nova_en.html

products that serve the Church's mission. Pope Leo addressed the gathering noting that technological innovation can be a form of participation in the divine act of creation and as such carries and ethical and spiritual weight, for every design choice expresses a vision of humanity: "the Church therefore calls all builders of AI to cultivate moral discernment as a fundamental part of their work to develop systems that reflect justice, solidarity, and a genuine reverence for life".[21] The Holy Father emphasized that interdisciplinary collaboration "embodies the dialogue between faith and reason renewed in the digital epoch and affirming that intelligence—whether artificial or human—finds its fullest meaning in love, freedom and relationship with God".[22] Theological and ethical insights reveal that AI as a positive tool with adequate safeguards can assist the world's populations. Yet space for reflection is necessary in the development of these tools, in order that the intrinsic worth value and dignity of the human person is upheld in any of our moves forward with artificial intelligence.

4. Reducing Economic Disparities through AI Driven Economy (Jonathan Warner)

4.1. The Coming of AI

The spread of AI is presenting new social and economic challenges. The economic benefits are (fairly) clear: higher productivity through the automating of some jobs, easier communication and access to knowledge, and the potential to help the poor and marginalized. Advances in AI reading of medical scans are improving diagnoses and detecting diseases and abnormalities earlier. The development of new ways to address some of the causes and effects of inequality (improving mobility, better and faster decision-making by government, for example) hold out the prospect of greater flourishing. *Emotional* support through chatbots is potentially also a life-enhancing use of AI, although it is not risk-free.[23]

One of the biggest fears of the increasing adoption of AI is concern over the loss of jobs. Over the past 50 years, the development of robotics has meant that many processes formerly done by people have been outsourced to machines. In general, these changes affected mainly relatively unskilled jobs, but more recent developments in AI have threatened jobs in other

[21] Pope Leo XIV, "Message of the Holy Father to the Participants in the Builders AI Forum" (6-7 November 2025), accessed 14 May 2026: https://www.vatican.va/content/leo-xiv/en/messages/pont-messages/2025/documents/20251103-messaggio-builders-aiforum.html. He also stated that the work of AI needs to be a profoundly ecclesial endeavour: "Whether designing algorithms for Catholic education, tools for compassionate healthcare, or creative platforms that tell the Christian story with truth and beauty, each participant contributes to a shared mission: to place technology at the service of evangelization and the integral development of every person".

[22] Pope Leo XIV, "Message AI Forum", 2025.

[23] The literature on the use of chatbots for mental and emotional health has been growing rapidly in recent years, especially since the advent of generative AI. For a recent review of the issues involved see Riccardo Volpato, *et al*., "Trusting Emotional Support from Generative Artificial Intelligence: A Conceptual Review", *Computers in Human Behavior: Artificial Humans,* vol 5 (2025), https://pmc.ncbi.nlm.nih.gov/articles/PMC11010238/pdf/nihms-1978097.pdf; for a review of empirical studies Young-Min Cho, *et al*. "An Integrative Survey on Mental Health Conversational Agents to Bridge Computer Science and Medical Perspectives", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023), accessed 29 April 2026: https://pmc.ncbi.nlm.nih.gov/articles/PMC11010238/pdf/nihms-1978097.pdf

sectors. Those without transferable skills of some kind, often the marginalized and disabled, look likely to be differentially adversely affected. In the past few years, concerns that the displacement of labour will lead to mass unemployment (and the social and psychological problems associated with it) have grown. Changes in work patterns are not new, though. The enclosure movement and industrial revolution of the late 18th and early 19th centuries saw massive labour market shifts and movements of population as people left farming areas to work in the newly developing factories. Skilled artisans in the textile industries found themselves out of a job. This led to social discontent, notably the activities of the machine-breaking Luddites.

Similarly, the closure of heavy industry and coalmines in the mid/late 20th century caused similar disruption and protests. Displaced workers could not find other jobs: a general recession in the economy, combined with their lack of suitable skills and of alternative employment left many without work and reliant on state benefits. Will the undoubted disruptions of the AI revolution have similar effects? Will skilled workers find themselves out of a job, with little prospect of finding another one?

There are reasons for optimism. Firstly, the benefits of automation and robotics have so far appeared relatively slowly, allowing more time for transition. In addition, if the advances are labour-enhancing, making workers more productive by taking over parts of their jobs, then unemployment may be avoided. Also, AI can open opportunities that increases the demand for certain kinds of work. For example, voice-sampling has made accent-changing possible: the customer service rep you call can now speak to you in a version of English comprehensible to you.

Secondly, if technology is labour-substituting (replacing workers, rather than upskilling them), new opportunities are likely to arise elsewhere, but probably in ways hard to predict. Today, banking is undergoing a major shift, as branches are being closed at a rapid rate. This causes disruption: to continue to work in the industry might require moving to a new area, which is difficult. But the skills of bank workers are more easily transferable to other tasks. Greater educational achievements also open up more opportunities. Learning how to teach yourself new skills is particularly valuable, especially as skills might be quickly outdated.[24]

Automation hasn't affected production in the ways initially thought. Robotics has produced some great advantages (Fiat's "handmade by robots" cars, precision engineering and key-hole surgery, for example) but progress in other areas has been slower. Even though using AI-driven robotic workers might be feasible, it is not necessarily economically viable Where visual inspection skills are important (such as checking ingredients for baking), using human workers continues to be a cheaper option,[25] although AI is now better at reading at least some medical

[24] See also David H. Autor, "Why are there still so many jobs? The History and Future of Workplace Automation", J*ournal of Economic Perspectives* 29 no. 3 (2015): 3-30; and the more recent utopian essay by Dario Amodei, "Machines of Loving Grace" (2024) accessed 18 November 2025: https://www.darioamodei.com/essay/machines-of-loving-grace

[25] Maja S. Svanberg, Wensu Li, Martin Fleming, Brian C. Goehring and Neil C. Thompson, "Beyond AI Exposure: Which Tasks are Cost-Effective to Automate with Computer Vision?", *MIT Working Paper*, https://futuretech-site.s3.us-east-2.amazonaws.com/2024-01-18+Beyond_AI_Exposure.pdf?ascsubtag=0000FT0000309530D1440077920240123212316.

images and scans than skilled doctors. This suggests that AI introduction will be sector-specific, meaning that the overall effects on jobs will come about relatively slowly.

The rollout of digitization with the arrival of the worldwide web in the 1990s provides a parallel: the promised boost in productivity materialized only later, and there was time to adjust to the changes it wrought. But in some areas the speed of transition is very quick: AI can now write computer code more quickly (and better code) than human coders. Prompt engineering jobs, which flourished when AI chatbots became widely available, are now in decline, as most people have learned how to write their own prompts. On the other hand, as populations age, there is an increasing demand for social care. It is unclear if this demand can be met by machines: for physical tasks, people still have an advantage. With mental/emotional help, chatbots have been available for several decades (since Joseph Weizenbaum's ELIZA[26]); but have considerable drawbacks. Being too affirming has led to chatbots advising clients to end their own lives; being less affirming leaves clients with emotional needs unmet. At present, the trend is to allow chatbots to refuse to answer, and training them to say: "I don't know" (which also helps deal with the problem of hallucinations).

4.2. A World with Less Work?

There is an important point missing from this analysis: the assumption that displacing humans from paid employment is a bad thing. If productivity growth continues, then less work is necessary, allowing for a shorter work week, or longer holidays, or earlier retirement, or some combination of the two.

Would eliminating the need for many kinds of toilsome labour be desirable, by allowing us to live more fulfilled lives—the utopian visions of Keynes[27] and Marx[28] finally realized? Many types of work give meaning to life; especially when work is a communal activity and so helps meet our needs for companionship and belonging. If work is done by machines, what can replace these benefits of it?

Firstly, work isn't limited to paid employment. Caring for children, the old and the sick can be fulfilling, too. Volunteering, community service and so on can provide fulfilment and joy as well. Some kind of way of recognizing the value of these activities might make it seem a more desirable option. Secondly, hobbies, sports and other leisure-time activities can also generate meaning, when there is appropriate interaction with others. Unfortunately, the arrival of smart phones has dealt another blow to community: loneliness seems now to be epidemic. For the young, today's digital natives, scrolling through endless Facebook posts and Tik Tok videos creates the impression that everyone has a better life. Research shows a correlation between owning a smart phone at an early age and mental illness, depression (and even suicide)

[26] Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: W. H. Freeman, 1976).

[27] Keynes, John Maynard "Economic Possibilities for our Grandchildren," in *Essays in Persuasion* (Harcourt Brace, 1932), 358-373

[28] Karl Marx and Frederich Engels, *Die Deutsche Ideologie* (1932), engl. tr. *The German Ideology* (Progress Publishers, 1964), 53: "[Communist society] makes it possible for me to do one thing today and another tomorrow, to hunt in the morning, to fish in the afternoon, rear cattle in the evening, criticise after dinner, just as I have in mind, without ever becoming hunter, fisherman, shepherd or critic".

a decade or so later.[29] Education for leisure needs to be part of the curriculum.[30] Therapy chatbots might help but cannot completely replace in-person meetings and activities—or hanging out with friends.

If we are to live lives free of toilsome labour, however, how shall we be able to obtain what we need? One solution would be to have a universal basic income (UBI)—pensions for all, perhaps. The profits made by businesses as a result of using machines rather than human labour would be taxed to provide the funds for this. At present, although UBI would appear to be a disincentive to work, evidence suggests that any such effect is small.[31] In any case, in a world where paid work is no longer seen as necessary for everyone, any disincentive effect would not be a problem. As we transition to this type of economy, those roles that are currently not amenable to replacement by AI, could continue to be paid, giving workers extra income on top of their UBI.

4.3. Regulation for Good

With the growth of AI, and especially with generative LLMs and text-to-graphics interfaces capable to create of deepfake images, the need for considering what regulation is necessary has increased. While the EU, the USA and other countries have introduced standards, the UK has decided on a softly- softly approach to regulation in the future, but for now the ability to innovate is to be encouraged. This doesn't mean no controls, firms may choose to self-regulate; and there are other standards already in place (on maintaining the privacy of users' personal data, and on age verification to prevent children from accessing inappropriate content, for example).

Regulatory problems aren't entirely new: industries which are highly regulated, or seen as unethical, have incentives to innovate when normal channels and procedures are blocked by regulation. The growth of the internet twenty years ago enabled new options for industries such as pornography and gambling. The arrival of generative AI has thrown up new opportunities and the need to amend regulation to catch up. What should be done about the rise of *deep fake* pornographic videos, and scams which are enabled by these and other tools? Unfortunately, regulation is often seen as violating privacy and is generally not enforceable extraterritorially—fakers and scammers can relocate to more accommodating jurisdictions, or just ignore the law. Consumers can also find ways around bans: when age verification for use of adult content sites came into force in Britain, the number of visits to the affected sites fell, but at the same time there was a rise in the number of downloads of VPN applications, which allow their users to appear as if they are somewhere else in the world.[32]

---

[29] Jonathan Haidt, "Get Phones Out of Schools Now", *The Atlantic*, 6 June 2003, is following in the footsteps of Neil Postman, *Amusing Ourselves to Death* (Methuen, 1985).

[30] This is perhaps a return to an Aristotelian approach to education, an education for a good life (Volf) rather than for work. Soft skills are now seen as at least equally important as traditional work-mandated skills.

[31] Evelyn Forget, "The Town with No Poverty: The Health Effects of a Canadian Guaranteed Annual Income Field Experiment", *Canadian Public Policy,* Vol. 37 No. 3 September 2011, 283-305.

[32] Liv McMahon, "VPNs Top Download Charts as Age Verification Law Kicks in", BBC News, accessed 14 May 2026: https://www.bbc.co.uk/news/articles/cn72ydj70g5o

There are other problems with regulating too soon. It increases the risk of industry concentration and monopolization, as large firms can more easily adapt to the regulation, and influence the shape it takes. Age verification via a third party would require the major players to co-operate to build the system, thereby increasing their power (as well as the risk of data leaks). Conversely, not all deep fake videos are necessarily undesirable: how could the traditional of political satire, for example, be appropriately permitted? Regulation needs to be carefully drawn, to allow for exceptions of this type. The balancing of uses of this kind against the horrors of harmful faked pictures is a challenge, just one of many that will need to be addressed over the coming decade.

5. Promoting Equitable AI in Healthcare: Design AI to Protects Human Dignity and Autonomy Starting from the most Vulnerable (Gianmarco Contino)

The integration of AI into healthcare is frequently described as transformative. AI systems now exploit large-scale clinical and biological data, from electronic health records and imaging repositories to genomic and wearables datasets, to support diagnosis, prognosis and treatment selection.[33] The anticipated gains are substantial: improved accuracy, efficiency and timeliness of care, and in some settings the possibility of extending high-quality expertise to resource-constrained environments.[34]

Yet this promise is shadowed by the risk that AI will entrench or even magnify existing inequities. A widely cited example is the commercial population health management algorithm in which racial bias emerged because the system learned to predict future healthcare costs rather than actual health needs. Since Black patients typically incurred lower costs for the same burden of illness, the algorithm systematically underestimated their risk and deprioritised them for additional care.[35] This is not an isolated technical glitch but a structural problem: when historical data encode inequitable access, algorithms that optimise on those data risk reproducing injustice at scale.[36]

5.1. Algorithmic Bias, Opacity, and the Fragility of Trust

Identifying and correcting such biases is essential but far from straightforward. Many clinically deployed models are proprietary and inaccessible to independent scrutiny.[37] Even when models are available, high-capacity models with limited explainability, whether due to technical complexity or institutional opacity, continue to challenge meaningful explanation and accountability. In clinical practice, this opacity directly affects patient autonomy and shared

[33] F. Jiang, et al. "Artificial Intelligence in Healthcare: Past, Present and Future". *Stroke and Vascular Neurology*, 2(4), (2017); G. Briganti, O. Le Moine. "Artificial Intelligence in Medicine: Today and Tomorrow", *Stroke and Vascular Neurology,* 2(4), (2017), doi: 10.3389/fmed.2020.00027.
[34] Eric J. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence", *Nature Medicine*, 25(1), (2019): 44-56.
[35] Z. Obermeyer, B. Powers, *et al*. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations", *Science*, 366(6464), (2019): 447-453.
[36] Crawford, *Atlas of AI*, 2021.
[37] Danton S. Char, et al., "Implementing Machine Learning in Health Care—Addressing Ethical Challenges", *New England Journal of Medicine*, 378(11), (2018): 981–983.

decision-making: if clinicians cannot understand the basis for an AI recommendation, they are poorly positioned to contest, contextualise or communicate it and patients themselves are unable to meaningfully assess, question or consent to the use of such systems in their care.[38]

This fragility of trust is not only technical but relational: both contribute to the fragility of trust. Viewing AI as an infallible *oracle* that provides the correct answer risks fostering excessive dependence. This epistemic outsourcing can undermine clinical judgment and restrict the consideration of patient values and preferences.[39] Conversely, if AI systems are experienced as inscrutable, biased, or misaligned with patient interests, they may erode confidence not only in technology but in the institutions that deploy it.

5.2. Uncertainty, Probabilistic Medicine and the Limits of Prediction

All medical decision-making already takes place under uncertainty. AI systems, which almost always output probabilistic risk estimates, inherit and sometimes amplify that uncertainty. Data limitations, label noise, distribution shifts, and biological variability all constrain what can be reliably predicted.[40] In the case of early detection, efforts to maximise sensitivity inevitably reduce specificity, increasing false positives and lowering the positive predictive value (PPV), particularly in low-prevalence settings. In oncology and cardiometabolic screening, such shifts can lead to substantial overdiagnosis and expose patients to unnecessary imaging, biopsies or treatments. These burdens are distributed unevenly across populations: groups already subject to fragmented care or limited follow-up capacity experience higher downstream harm, while algorithmic thresholds tuned on majority cohorts often perform worse in minority groups, amplifying disparities in accuracy, overtreatment and access to confirmatory diagnostics.

Responding to this reality requires to move away from the persistent misconception that AI systems deliver better-than-human or fully objective medical judgments toward explicit handling of uncertainty. Technical strategies such as ensemble learning, calibration and uncertainty quantification can help characterise when a model is confident and when it is not.[41] At the same time, continuous learning systems and rigorous post-deployment monitoring are needed to ensure that model performance remains acceptable as clinical practice, patient populations and data infrastructures evolve.[42]

Uncertainty is a profound ethical issue, not just statistical, that links to responsibility and justice, often harming the vulnerable and future generations. Current regulatory bodies explicitly recognise this. The World Health Organization (WHO)[43] cautions that overreliance on AI systems may "undermine human autonomy, obscure accountability, and exacerbate existing inequities" if uncertainties are not transparently communicated. The US Food and

[38] Thomas Grote, Philipp Berens, "On the Ethics of Algorithmic Decision-Making in Healthcare", *Journal of Medical Ethics*, 46(3), (2020): 205-211.
[39] Grote, "Ethics of Algorithmic", 2020.
[40] Jason Wiens, *et al*. "No Free Lunch for Early Detection of Disease", *The Lancet Digital Health*, 1(5), (2019): e232-e233.
[41] E. Begoli, et al. "The Need for Uncertainty Quantification in Machine-Assisted Medical Decision-Making", *Nature Machine Intelligence*, 1(1), (2019): 20-23.
[42] Alvin Rajkomar, *et al*. "Machine Learning in Medicine", The New England Journal of Medicine, 380(14), (2019): 1347-1358.
[43] World Health Organization (WHO), *Ethics and governance of artificial intelligence for health.* WHO (2021).

Drug Administration (FDA)[44], through its Good Machine Learning Practice (GMLP) guidance, requires traceability, clarity of intended use, and ongoing performance monitoring to prevent AI systems from drifting into unsafe or misleading behaviour. Similarly, the European Medicines Agency (EMA)[45] emphasises the need for explainability proportional to clinical risk and mandates rigorous lifecycle oversight for AI-enabled medical devices. These regulatory positions highlight the ethical stakes of uncertainty: not merely whether the model is correct, but whether its limitations are communicated in ways that allow patients to make informed choices and clinicians to retain professional responsibility. How uncertainty is communicated, who is told what, in which terms, and with what options, shapes the autonomy and psychological burden of patients and clinicians alike. Patients often misinterpret AI performance, frequently assuming that algorithmic outputs are objective, error-free, or universally applicable. Studies in diagnostic AI show that patients tend to equate the use of technology with greater accuracy and impartiality, a form of automation bias that increases willingness to accept recommendations without question.[46] This effect is more pronounced among individuals with lower health literacy or limited familiarity with probabilistic reasoning. Conversely, when errors occur, trust can collapse abruptly, especially in populations historically underserved or misdiagnosed by healthcare systems. These dynamics underscore the need for patient-facing explanations that clarify what an AI system can and cannot reliably do, how uncertainty affects risk, and why clinical judgment remains essential in contextualising algorithmic outputs.

5.3. Embedding AI within Patient-Clinician Decision-Making

Given these limitations, AI in medicine should be framed as a decision support technology rather than a decision replacement technology. Integrating AI into clinical workflows in ways that respect patient autonomy requires transparent communication of its capabilities, limitations, and potential sources of error.[47]

This implies at least three practical commitments: 1) Explainability appropriate to context: not every model needs to be interpretable in the same way, but clinicians require enough insight to judge whether the output is plausible in light of the patient's history and values, and to explain their reasoning to patients, 2) Retention of professional responsibility: AI recommendations must remain contestable. Legal, institutional and professional norms should clarify that responsibility for clinical decisions cannot be delegated entirely to algorithms; 3) Shared decision-making that includes AI as one voice among many: patients should be informed when AI is used in their care and, where feasible, have opportunities to question or decline its use, particularly in high-stakes settings or where personal or ethical values are deeply implicated.

---

[44] U.S. Food and Drug Administration (FDA). *Good Machine Learning Practice (GMLP) for Medical Device Development: Guiding Principles.* FDA/Health Canada/MHRA Joint Document. (2021).

[45] European Medicines Agency (EMA). *Reflection Paper on the Use of Artificial Intelligence in the Medicinal Product Lifecycle.* (2023): EMA/226338/2023.

[46] Christopher Robertson, *et al*. "Diverse Patients' Attitudes towards Artificial Intelligence (AI) in Diagnosis", *PLOS Digital Health* 2(5) (2023): e0000237. https://doi.org/10.1371/journal.pdig.0000237

[47] Effy Vayena, *et al*. "Machine Learning in Medicine: Addressing Ethical Challenges", *PLOS Medicine*, 15(11), (2018): e1002689.

5.4. Designing for Equity: From Majority Norms to Plural Values

A central challenge for equitable AI in healthcare is alignment with the diverse values, experiences and bodies of patients in multicultural societies. Training datasets often over-represent majority populations or those with greatest access to healthcare systems. This can produce models that perform well for "typical" cases while systematically misclassifying or underserving minority groups, people with rare diseases, or those with intersecting vulnerabilities.[48]

Avoiding this "normalisation" of the majority requires both technical and institutional strategies: 1) Representative and reflexive datasets: investment in diversifying training data, and in systematically documenting where data are missing or biased, is essential. Dataset documentation and model cards can make these limitations visible and auditable;[49] 2) Participatory and value-sensitive design: Engaging patients, carers, frontline clinicians, and marginalised communities in the design, evaluation and governance of AI systems can surface value conflicts and context-specific harms that may be invisible to developers; 3) Regulation and governance: Principles, such as fairness, accountability, transparency, privacy, are a necessary starting point but insufficient on their own.[50] They must be connected to concrete mechanisms: impact assessments, independent auditing, clear routes for contestation and redress, and sanctions where harms occur.

These measures are not merely procedural. They embody a normative stance: that AI in healthcare must serve the most vulnerable and not only the already well-served.

5.5. Towards Equitable and Humane AI in Medicine

AI in medicine sits at the intersection of extraordinary technical capability and profound human fragility. It can help detect disease earlier, personalise treatment, and extend scarce expertise but it can also encode structural injustice, obscure responsibility, and narrow the space for human judgment and patient self-determination.

Promoting equitable AI in healthcare therefore requires more than technical excellence. It demands continuous critical scrutiny of data and models; institutional commitments to transparency, accountability and participation; and regulatory frameworks that prioritise those who are most at risk of exclusion or harm. Above all, it requires that AI be explicitly situated within a relational model of care, where technology supports rather than supplants the moral responsibilities of clinicians and the autonomy of patients. Only under these conditions can AI contribute not merely to high-performance medicine, but to a form of care that honours the dignity, diversity and vulnerability of those it seeks to serve.

---

[48] Buolamwini, and Gebru, "Gender Shades", 2018; Adam S. Adamson, and Adewole Smith, "Machine Learning and Health Care Disparities in Dermatology", *JAMA Dermatology*, 156(11), (2020): 1127-1128; Raji and Buolamwini, "Actionable Auditing", 2019.

[49] Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave, *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research,* London: Nuffield Foundation, 2019.

[50] Floridi, "A Unified Framework", 2019.

5. Conclusion

The concern of each author has been to utilise their area of expertise in such a way that they identify key issues within their own discipline, while also having a common concern for the flourishing of the human person. We hope that this paper will stimulate the reader's own creative imagination and if we have raised questions concerning the developments in AI, then we shall have achieved our aim. The movement forward in this area is inevitably transdisciplinary and in that dynamic, we may assist one another, both to benefit from the extraordinary possibilities in AI and to be reflectively aware of key challenges and concerns for the fragile human persons we continue to be.