Routledge Advances in Social Work

AI AND THE DISRUPTION OF THE SOCIAL

CHALLENGES FOR SOCIAL WORK EDUCATION AND PRACTICE

Edited by
Goetz Ottmann and Carolyn Noble

ROUTLEDGE

AI and the Disruption of the Social

AI and the Disruption of the Social analyses digital automation and the disruption of both the social and social work, in response to rapid advances in digital technology, automated decision-making, and artificial intelligence (AI) within systems of governance, work, education, and everyday life.

Establishing that we're now entering a 'fourth industrial revolution' driven by late capitalism, this volume approaches technological advancements with a critical lens, exploring how digital automation is disrupting the social, and particularly the lives of the most vulnerable. Contributors suggest that AI systems are amplifying existing socio-economic, gender, and racial biases and stereotypes in our society, suggesting that an ethical and progressive political response is needed to counteract this. Arguing that digital advancements represent a push to commodify human 'intelligence' while advancing state surveillance projects and coercive control, this work explores how automation is transforming the knowledge and skills required to thrive in human services roles and within a civil society.

This book highlights important philosophical, epistemological, ethical, educational, and practice challenges for governments, human services practitioners, and educators and will be of interest to all scholars and students of social work, social policy, and sociology.

Goetz Ottmann is a Senior Lecturer at Federation University in Melbourne, Australia. He is the author of several books, book chapters, and numerous peer-reviewed journal articles. His research interests include critical social theory, authoritarian 'welfare', social policy, and aged and disability care. His latest co-authored books include *The Challenge of Right-wing Nationalist Populism for Social Work* (2020), *Post-Pandemic Welfare and Social Work: Reimagining the 'New Normal'* (2024), and *AI and the Disruption of Welfare* (2026), published by Routledge.

Carolyn Noble is a Professor Emerita at ACAP, Sydney, and Victoria University, Melbourne, Australia. She is the author of several books, chapters, and peer-reviewed articles. Her latest co-authored and co-edited books include *The Challenge of Right-wing Nationalist Populism for Social Work* (2020), *Post-Pandemic Welfare and Social Work: Re-imaging the 'New Normal'* (2024), *The Routledge International Handbook of Feminisms in Social Work* (2024), and *AI and the Disruption of Welfare* (2026). She is the current editor-in-chief of IASSW Social Dialogue: https://socialdialogue.online/.

Routledge Advances in Social Work

Social Work with Unmarried Mothers and Their Children
Learning from the Past
Viviene E. Cree and Robert MacKenzie with Douglas Edgar and Rose Foy

Radical Hope in Social Work
Edited by Carole Zufferey and Chris Horsell

Social Work, the Americanization Movement and the Construction of Americans 1874–1930
Yoosun Park and Michael Reisch

Community Social Labs and Social Work
Bridging Gaps Between Universities and Communities to Solve Social Problems
Edited by Janestic Mwende Twikirize and Ann Christin E. Nilsen

AI and the Disruption of the Social
Challenges for Social Work Education and Practice
Edited by Goetz Ottmann and Carolyn Noble

For more information on the series and a full list of titles, please visit
https://www.routledge.com/Routledge-Advances-in-Social-Work/book-series/RASW

AI and the Disruption of the Social

Challenges for Social Work Education and Practice

Edited by Goetz Ottmann and Carolyn Noble

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2027
by Routledge
4 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2027 selection and editorial matter, Goetz Ottmann and Carolyn Noble;
individual chapters, the contributors

The right of Goetz Ottmann and Carolyn Noble to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

The electronic version of this book was funded to publish Open Access through Taylor & Francis' Pledge to Open, a collaborative funding open access books initiative. The full list of pledging institutions can be found on the Taylor & Francis Pledge to Open webpage.

The Open Access version of this book, available at www.taylorfrancis.com, has been made available under a Creative Commons Attribution-Non Commercial-No Derivatives (CC BY-NC-ND) 4.0 International license.

Any third party material in this book is not included in the OA Creative Commons license, unless indicated otherwise in a credit line to the material. Please direct any permissions enquiries to the original rightsholder.

For Product Safety Concerns and Information please contact our EU representative GPSR@taylorandfrancis.com. Taylor & Francis Verlag GmbH, Kaufingerstraße 24, 80331 München, Germany.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN: 9781041041955 (hbk)
ISBN: 9781041043423 (pbk)
ISBN: 9781003627920 (ebk)

DOI: 10.4324/9781003627920

Typeset in Times New Roman
by codeMantra

Contents

List of Tables *viii*
List of Contributors *ix*
Acknowledgements *xvi*

1 AI and the Disruption of the Social: An Introduction 1
GOETZ OTTMANN AND CAROLYN NOBLE

2 Epistemic Agency and the Politics of Social Work 14
GOETZ OTTMANN

3 Social Work in the Hypermodern: Advancing Humanism in the Age of Smart Machines 27
DAVID P. MOXLEY

4 Can Artificial Intelligence Be Controlled? A Social Science Perspective 46
HERMANN ASTLEITNER

5 The Politics of Information and Communication Technologies (ICTs), Artificial Intelligence (AI), and Implications for Social Work, Justice, and Advocacy 60
CAROLYN NOBLE

6 Critical Scholarship in the Age of Large Language Models 72
RUSSELL SHUTTLEWORTH

7 AI and the Corporate University: Constructing New Imaginaries or Reinscribing Old Stratifications? 87
SHARON ARIS AND GEORGIA VAN TOORN

8 AI and Indigenous Knowledge Holders: When Culture Becomes Security: Cultural Security and Epistemic Sovereignty in the Sámi Context 99
BARBARA SCHABOWSKA

9 Masculinism and Gender Binarism in AI: How Artificial Intelligence Reproduces Patriarchy in the Digital World 111
BOB PEASE

10 The Helping Relationship in New Online Relational Spaces: AI Chatbots and the Reconfiguration of Trust in Social Work 126
LUCA PAVANI AND MARA SANFELICI

11 Care, Artificial Intelligence (AI), and Ethics: New Opportunities and Challenges for Social Workers 139
ELENI PAPOULI

12 Bridging Worlds: Artificial Intelligence, ChatGPT, and Social-Work Ethics 153
ALBERTO BALLESTERO, JUAN-JESÚS VISCARRET, AND MARÍA-JESÚS ÚRIZ

13 A Cruel Kindness? The Limitations of Logic-Based Artificial Intelligence in Navigating Ethical Dilemmas within the Human Services Context 169
CLEMENT CHIHOTA AND TARIRO CHIHOTA

14 Social Work Ethics and Generative AI 177
MICHAEL BALKOW

15 Virtual Encounters in Bereavement Care: Ethical and Clinical Perspectives on AI and Virtual Reality 190
SEOUNGHOON OH, RANAA ALMASOUDI, AND CATRIONA MAYLAND

16 Bias in the Machine: How Artificial Intelligence Perpetuates Ableism in Society 206
SIMONA AGINSKAITĖ AND EGLĖ ŠUMSKIENĖ

17 Artificial Intelligence between the Lighthouse and the Storm: Potentials and Risks for Social Work 219
RENATA NUNES, MELISA CAMPANA ALABARCE, DAVID ALONSO GONZÁLEZ, AND ANDRÉS ARIAS ASTRAY

18 The Use of AI-Based Counselling: Professional and Ethical Implications for Social Work 233
MONIKA ALAMDAR-NIEMANN AND SABRINA HEINL-VAKO

19 What's Wrong with AI as a Human Rights Teacher for Social Workers? Reflections on a Teaching Innovation Project in Spain 245
MARÍA INÉS MARTÍNEZ HERRERO, FRANCISCO J. LORENZO-GILSANZ, AND SERGIO BARCIELA FERNÁNDEZ

Index *263*

Tables

15.1 Case summaries of VR bereavement encounters in Meeting You 193
19.1 Summary of AI chat development instructions 249
19.2 Overview of AI chatbot project outcomes 251

Contributors

Simona Aginskaitė is a PhD candidate in Social Sciences at the Faculty of Philosophy, Vilnius University. In her doctoral research, she examines how technological developments reshape societal expectations regarding the abilities of persons with disabilities and how these shifts contribute to the transformation of contemporary understandings of disability. Simona Aginskaitė also conducts monitoring and research activities within the disability rights advocacy organisation Lithuanian Disability Forum. Her research interests include the representation of persons with disabilities in the media, violence against women with disabilities, and the reproductive rights of women with disabilities. She has also studied journalism and leadership. Her professional experience includes work at the Lithuanian National Radio and Television, the Agency for the Protection of the Rights of Persons with Disabilities under the Ministry of Social Security and Labour, and long-term internships in disability organisations in France and the United States. She is also an alumna of the U.S. Department of State's Community Solutions Program.

Monika Alamdar-Niemann is Professor Emerita of Social Work at IU International University of Applied Sciences, Germany. From 2018 to 2024, she served as Professor of Social Work, with teaching and research interests in social work methods, integrative methodology, migration, intercultural pedagogy, gender perspectives, and education in international contexts. She has held academic appointments at several German universities, including the Free University of Berlin and the University of Hamburg, and has extensive experience in international higher education, including the Erasmus programme. She received her doctorate in education from the Free University of Berlin. As part of her scholarly work, she has published research studies and expert reports on the educational trajectories of children from migrant families, parenting practices, and policy-related expertise for the German federal government. Her more recent publications focus on integrative methodological approaches, particularly in relation to mothers with refugee experience, offering significant contributions to the evaluation and advancement of this concept.

Ranaa Almasoudi is a Lecturer at the University of Tabuk and a Nurse Researcher at the University of Sheffield and the University of Tabuk. She is currently working in the field of virtual reality and digital health.

David Alonso González is a Full Professor at the University Department of Social Work and Social Services at the Faculty of Social Work (Complutense University-UCM). He has been Professor of Social Work since 1999 (1999–2007: UCLM and 2007–present: UCM). He holds a Diploma in Social Work, a Bachelor in Social and Cultural Anthropology, and a DEA in Sociology, and is a Doctor in Social Work (Extraordinary Award). David is a member of the UCM Research Group 'Psychosocial Factors and Social Intervention', a member of the Research Group in Social and Educational Gerontology (UCLM), and a member of the Institute of Knowledge Technology (www.ucm.es/ITC), and has been recognised with two six-year research periods (2022). He has experience in University Management as Vice-Dean at the Faculty of Social Work and as Coordinator of Innovation and Digitalisation at the General Foundation of UCM for 11 years. He is also coordinator at UCM for European Projects. His research interests include social innovation and social work, technology and social work, social work with groups, international social work, online pedagogy, active ageing, and the social and educational work.

Andrés Arias Astray is Chair of Social Work and Social Services at the Complutense University of Madrid, co-director of the research group, with mention of excellence, Psychosocial Factors and Social Intervention, director of GrupoLab, and a member of the Institute of Knowledge Technology (ITC). His research interests focus on issues related to social intervention and, particularly, its critical analysis through the articulation of a social psychology of techno-scientific knowledge, reflection on human nature in relation to the exercise of social professions, work with groups, the elderly, innovation, gender violence, and the development of instruments for the evaluation and orientation of professional practice in the social field.

Sharon Aris is a Nexus fellow and Lecturer in the School of Social Sciences at the University of New South Wales and a co-convenor of the Australian Sociological Association Sociology of Education Special Interest Group. Her research focuses on hidden curriculums, how different knowledges are produced and transmitted in secondary and tertiary education and through families, and how education can be made more genuinely inclusive.

Hermann Astleitner is Associate Professor of Education at the Paris Lodron University of Salzburg (Austria). His research is on instructional systems design, human development, and education, as well as social research methods and theory building. His recent publications include Astleitner, H., Bains, A., & Hörmann, S. (2023). The effects of personality and social media experiences on mental health: Examining the mediating role of fear of missing out, ghosting, and vaguebooking; Astleitner, H. (2024). We have big data, but do we need big theory? Review-based remarks on an emerging problem in the social sciences;

and Astleitner, H., & Schlick, S. (2024). The social media use of college students: Exploring identity development, learning support, and parallel use.

Michael Balkow is a Senior Lecturer in Social Work at the University of Derby, holding an MA in Social Work. He is currently leading modules on the MA, BA and Apprenticeship Social Work degrees. His research interests include transgender people and social work, children's mental health, social work writing and bureaucracy, the use of philosophy in social work, ethics and assisted dying in terminally ill people, AI and social work practice, AI Chatbots and mental health services, and freedom of speech in academia. He is a registered social worker, with a professional background predominantly in children's social work, across child protection, children with disabilities, children with mental health issues, and public and private law family proceedings.

Alberto Ballestero holds a Bachelor Degree in Philology from the National University of Distance Education (Spain), a Master's Degree in Language, Immigration and Social Work from the University of Lleida, and a PhD in Social Work Ethics from the Public University of Navarre (Spain). He is Associate Professor of Philosophy at the Public University of Navarre and researcher in the EFIMEC research group. His research interest focuses on social work ethics. Author of, and contributor to, numerous academic articles, books, and chapters, he has published his research in impact journals and participated in collective works as author.

Melisa Campana holds a Bachelor's and PhD degrees in Social Work (National University of Rosario, Argentina). She is an Associated Researcher at CONICET (Argentina) and Principal Professor at the National University of Rosario. She has led the Research Centre on Governmentality and State (CIGE-PEGUES) and has been editor-in-chief of *Cátedra Paralela* Journal. She is a member of the Social Work Action Network (SWAN) and the Work Group on Street-Level Bureaucracy (CLACSO). Currently, she is a contracted researcher at Complutense University of Madrid, conducting the Project Frontline Innovative Practices Bank, under Marie Sklodovska-Curie Cofund Program UNA4CAREER. Her research interests include critical social work, social policy, social assistance, social innovation, critical epistemologies, and decolonial feminisms.

Clement Chihota is a Senior Lecturer and the acting Head of Social Work, Community and Human Services at Federation University Australia. His professional career spans multiple countries and disciplines. Originally from Zimbabwe, he has served as a Subject Specialist in Applied Linguistics at the Zimbabwe Open University, an Academic Writing Consultant and lecturer in Postgraduate Development at the University of Cape Town, a Child Protection Social Worker with Child, Youth and Family Services (now Oranga Tamariki) in New Zealand, and a Social Work Educator at Bethlehem Tertiary Institute. Drawing on this rich and interdisciplinary background, Clement's teaching and research are grounded in student- centred pedagogy and a deep commitment to integrating diverse perspectives within social and community work education.

His work bridges practical frontline experience with scholarly inquiry, enabling nuanced explorations of contemporary issues—including the emerging question of whether artificial intelligence can independently make ethical decisions within the human services field.

Tariro Chihota holds a BA in Logic, Computation, and Linguistics and is currently studying towards a Bachelor of Health Sciences degree and pursuing a future career in Occupational Therapy. She has an academic background in computational logic and mathematics, which has strengthened her analytical reasoning and structured problem-solving abilities. With a growing interest in human services and health-related practice, Tariro is passionate about person-centred care and supporting individuals to enhance their well-being and everyday functioning. Her emerging work reflects a commitment to bridging technical, systems-oriented thinking with meaningful, human-focused outcomes—an approach that continues to shape her developing practice within the health sciences field.

Sergio Barciela Fernández holds a PhD in Migration and Development from Comillas Pontifical University (CPU), Madrid. He is a lecturer and researcher at CPU's Institute for Migration Studies, with advanced degrees from the Pontifical University of Salamanca and Universidad Nacional de Educación a Distancia (UNED). For over 12 years, he has coordinated migration programmes at Spanish Caritas.

Sabrina Heinl-Vako is a Research Associate at the International University of Applied Sciences, where she is affiliated with the distance learning programme in Social Work. In this role, she supports faculty members in teaching, the development of instructional materials, and the ongoing advancement of the degree programme. She holds both a Bachelor's and a Master's degree in Social Work and has gained professional experience in the fields of addiction services and integration support. Her academic interests focus in particular on political action within social work as well as on the significance of adverse childhood experiences for professionals working in child protection. As a board member of the Bavarian Association of Social Work Professionals, she is actively committed to improving working conditions and strengthening the professionalisation of social work.

Inés Martínez Herrero, PhD, is a Senior Lecturer at the Department of Social Work at UNED, Spain. She is a qualified social worker and holds a degree in Psychology. With teaching and research experience in England and Spain, her research focuses on human rights in social work, social work histories, and critical social work, including facing technological innovation.

Francisco J. Lorenzo Gilsanz has a PhD in Sociology and is a Social Worker. He is a Senior Lecturer in the Department of Social Work at UNED. He has coordinated multiple research projects on poverty, social exclusion, migrations, and social capital. As a Social Worker, he has extensive experience in the field of drug addiction and homelessness.

Catriona Mayland is a Senior Clinical Lecturer and Honorary Consultant in Palliative Medicine at the University of Sheffield. Her expertise lies in bereavement care, patient experience, and palliative care practice.

David P. Moxley is a Professor in the University of Alaska Anchorage College of Health and serves as Principal Investigator of the Alaska Interdisciplinary Distance Learning Consortium in Behavioral Health and the Alaska Careers in Behavioral Health.

Renata Nunes holds a Bachelor's and a PhD degree in Social Work from the Federal University of Santa Catarina (UFSC), Brazil, and Universidad Complutense de Madrid, Spain. She served as an Assistant Professor at the UFSC and was a member of two research groups dedicated to the study of the public social assistance system and social policies in Brazil. Currently, she is a fellow researcher at GrupoLab, a research group at UCM that focuses on Social Work and Social Intervention. Renata's research interests include critical social work, intervention in social work, social protection, and family studies from a gender perspective.

Seounghoon Oh is a researcher at Durham University, specialising in palliative social work with a focus on the integration of virtual reality (VR) in bereavement care. His doctoral research explored decision-making processes for children with life-limiting conditions, focusing on the intersection of bioethics, social justice, and palliative care policy. Currently, Seounghoon is involved in quantitative and qualitative research on generative AI at Durham University, exploring its implications for academic law degrees. His extensive teaching and conference presentations address critical issues in anti-oppressive health care and the role of social workers in disaster response and palliative care. Seounghoon's work reflects a commitment to improving palliative care practice through innovative technological integration, ensuring ethical standards, and promoting patient and family engagement in care processes.

Eleni Papouli is Associate Professor of Social Work Ethics and Creative Learning Approaches in the Department of Social Work at the University of West Attica (UNIWA). She is a qualified social worker with a background in drama therapy and professional experience in the child protection sector in Greece. She has more than 20 years of teaching experience in higher education and holds a teaching fellowship award from Munich University of Applied Sciences (MUAS), Germany (2015). She regularly participates as a guest speaker and expert masterclass (ECSWE 2023, Portugal) at international conferences and educational programmes (Sweden, Germany), and her academic work focuses on social work ethics and creative and arts-based learning methods. She is currently a member of the editorial board of *Social Work Education: The International Journal*. She is also an international PhD supervisor and academic Erasmus+ coordinator. Eleni is a founding member of the Interdisciplinary Laboratory for Research, Education, and Disability Support (DI.E.E.Y.P.A.) at UNIWA, and an affiliate member of Leeds Beckett University's Creative Arts and Health Research Laboratory (CAHREL). She is co-editor and writer of the

book *Creative Critical Pedagogies: Using Arts-Based Approaches for Online and In-Person Learning* (2024).

Luca Pavani is a postdoctoral researcher at the Department of Law, Political, Economic and Social Studies of the University of Eastern Piedmont, Alessandria (Italy). His research interests focus on eco-social approaches and sustainability in social work, and critical social work theory, with particular interest in anti-oppressive approaches and posthuman and neo-materialist philosophy.

Bob Pease is an Adjunct Professor in the School of Social Sciences at the University of Tasmania and a Honorary Professor in the School of Humanities and Social Sciences at Deakin University in Victoria, Australia. His most recent books are *Doing Critical Social Work* (co-editor, 2016), *Men, Masculinities and Disaster* (co-editor, Routledge, 2016), *Radicals in Australian Social Work* (co-editor, 2017), *Critical Ethics of Care in Social Work* (co-editor, Routledge 2018), *Facing Patriarchy: From a Violent Gender Order to a Culture of Peace* (2019), *Post-Anthropocentric Social Work: Critical Posthuman and New Materialist Perspectives* (co-editor, Routledge, 2021), *Undoing Privilege: Unearned Advantage and Systemic Injustice in an Unequal World* (2022), and *Posthumanism and the Man Question: Beyond Anthropocentric Masculinities* (Routledge, 2013).

Mara Sanfelici is an Associate Professor at the University of Milan Bicocca (Italy), where she teaches Social Work Methodology and Theories and Policy Practice in Social Work. Her main research interests relate to anti-oppressive practice in social work, digitalisation of social work, and social work with people in poverty.

Barbara Schabowska, PhD, is a social science researcher at the Korczaka University, Poland. She is a member of a research team that explores the potential applications of AI in psychological diagnostics. Her primary research focus lies in cultural security integrating cultural and social science perspectives with mental health research.

Russell Shuttleworth worked as a social worker and disability support worker for years. Trained as an anthropologist and social worker, he taught in these disciplines, as well as sexuality studies and disability studies at various universities including San Francisco State University, University of California Berkeley, University of Sydney, and Deakin University. As a researcher and scholar, he is internationally recognised in disability anthropology and critical disability studies. Russell has recently retired as a university academic and devotes his time to writing (applying critical theory and existential phenomenology to current topics), making music, and art. He is currently affiliated with Creative Imaginaries, Sydney, NSW.

Eglė Šumskienė is a social worker, sociologist, and expert on disability, mental health, and human rights issues. Since 2002, she works for Vilnius University Social Work and Social Welfare Department, and in 2022, she took the position

of a Professor. From 2018 to 2021, she was the Head of the Department of Social Work and Social Welfare, and from 2021 to 2024, she was the director of the Institute of Sociology and Social Work. From 2003, she actively works with the Lithuanian NGO sector, focusing on disability, mental health, and human rights. During this period as an expert, she was involved in various international and national projects focusing on such areas as human rights and disability, employment, discrimination, mental health care, policy, and social integration of people with psychosocial disabilities. E. Sumskiene is the editor of one monograph and the author of over 60 scientific publications covering issues of disability, human rights, and mental health.

María-Jesús Úriz holds a Bachelor's degree in Philosophy from the University of Navarre (Spain), a Bachelor's degree in Psychology, and a PhD in Philosophy from the University of the Basque Country (Spain). She is Full Professor of Philosophy at the Public University of Navarre and the director of EFIMEC research group. Her research interest focuses on social work ethics. Author of, and contributor to, numerous academic articles, books, and chapters, she has published her research in impact journals and participated in collective works as an author.

Georgia van Toorn is a Lecturer in the School of Social Sciences at the University of New South Wales and an Associate Investigator at the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S). Specialising in international social policy, politics, disablement, and social justice, she explores global transformations in welfare governance, with a particular focus on processes of marketisation, the commodification of social care, and the growing impact of data analytics and algorithmic decision-making in the public sector.

Juan-Jesús Viscarret holds a Bachelor's degree in Social Work from the University of Zaragoza (Spain), a Bachelor's degree in Political Science and Sociology from the Pontifical University of Salamanca, and a PhD in Sociology from the Public University of Navarre (Spain). He is a Senior Lecturer in Social Work and Social Services at the Public University of Navarre and a researcher in the EFIMEC research group. His research interest focuses on social work and youth. Author of, and contributor to, numerous academic articles, books, and chapters, he has published his research in impact journals and participated in collective works as an author.

Acknowledgements

We know that a book is nearing completion when we sit down, begin to reflect on the process, and gather a list of people who have made it possible. It is both a happy and sad time for us editors, as we say goodbye to the project that has taken a considerable amount of our individual and collective time, energy, thinking, and organising for the better part of a year.

This second book on AI and its impact on society's evolving relationship with technology and its influence on societies' epistemological integrity, governance systems, and professional practices has allowed us to provide further direction and potential suggestions to address some of the many issues the rapid rise and adoption of AI technologies across the world has presented. To construct this book, we invited comments, insights, and critical feedback from scholars and students alike.

The response has come from a dedicated group of academics and practitioners from across the globe who have presented their ideas, research, arguments, reflection, and scholarship to unpack and distil the many competing complexities surrounding AI and its potential to disrupt the social in ways that will deplete our communities and exacerbate disparities of knowledge, power, and wealth. For professionals working in the social and health services, AI is spelling major changes not only in the way clients are affected by the new technologies but also how professions such as social work are fundamentally reshaped by seemingly 'intelligent' machines.

We thank our co-authors for contributing to this book, adhering to our deadlines, and responding to our editorial suggestions despite considerable work commitments. We thank them all for their valuable insights and look forward to further collaborations.

Our thanks extend, as always, to the staff at Routledge that helped us get this book to fruition: Claire Jarvis and Sully Evans for their support and publishing guidelines and Sarah Lewis and her editorial team for pulling it all together. This is our fourth contribution to the very popular *Routledge Advances in Social Work Series*. We believe our contributors have offered valuable insights for your consideration.

Finally, we thank each other for a productive working relationship. As stated before, we share a love of critical curiosity and like to challenge each other to explore new and thought-provoking ideas. We share a passion for social work's

commitment to emancipatory practice and continue to use this lens to explore our profession's values, mission, ontology, and epistemology. As AI's disruption of the social is unfolding, it is vital that we become more aware of and learn to confront the challenges harboured by this rapidly evolving technology, so that in the not-too-distant future, our professions will be still professions and not simply interactive routines in automated processes.



Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1 AI and the Disruption of the Social
An Introduction

Goetz Ottmann and Carolyn Noble

Introduction

The impact of Artificial Intelligence (AI) on society is still unsettled; its future is yet to be fully determined. However, what is getting increasingly clear is that, given its current, business-defined trajectory, AI has the potential to significantly re-shape political economies, global and domestic hierarchies of power, and knowledge production; disenfranchise labour; polarise political processes; deplete cultural resources; re-shape societal structures; centralise education; transform gender; undermine the basis on which truth claims can be constructed and advanced; and weaken human agency (Acemoglu et al., 2025, 7 October; Acemoglu et al., 2025; Carvalho, 2026; Zhang et al., 2025). In other words, AI has the capacity to fundamentally disrupt the social. Yet current research on the impact of AI on social services professionals, vulnerable individuals, and communities remains scarce and is often drowned out by corporate and political memes that reiterate in new terms the old neoliberal adage that There Is No Alternative (TINA) – AI is essential for economic growth and productivity. Technological infrastructure and AI technologies are progressively embedded in the very fabric of lived experience, shaping the textures of human subjectivity, cultural identities, and personal life (Elliott, 2021). This is reshaping how we communicate, trust, relate, work, and make meaning of the world we inhabit, as it becomes increasingly difficult to extricate ourselves and our institutions from the influence and their destructive potential and bridge the widening knowledge gap that would allow us to manage our exposure to AI.

Disruptions of the social, with their capacity to dis-embed individuals from socio-economic structures, are often justified in terms of potential increases in productivity, reductions in labour costs, and the elimination of repetitive tasks, freeing up human labour to engage in other, more rewarding pursuits. However, there is rising concern even among the executives of tech corporations (in February 2026, Mrinank Sharma's public resignation from Anthropic and Jimmy Ba's from Grok AI revealed as much) that AI threatens human existence in a number of ways (Kohler, 2026, February 16). When digital automation becomes the primary focus of an unrestrained capitalism, human labour becomes optional. AI systems are rapidly increasing in complexity and sophistication (Barnes & Hutson, 2024). Ignoring or downplaying the problematic aspects of AI risks undermining

DOI: 10.4324/9781003627920-1
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

conventional structures of knowledge production and what counts as 'real', 'true', or 'human', raising questions about how we ought to understand our own agency (Elliott, 2021). Shipley and Williams (2023) argue that AI technologies are being relentlessly engaged in "subjugating every individual, not just particular groups, through complete and irreversible forms of objectification, alienation, and datafication" (p. 619) with unpredictable social consequences, many of which are yet to be identified and/or properly understood (Elliott, 2021).

AI digitises and monetises knowledge that is being acquired at an incredibly low cost. In fact, in many cases, it is simply extracted from open-access datasets, publicly available websites, academic research and publications, encyclopaedias, and social media sources – including the data collected through digital surveillance processes. The rollout of AI-based knowledge products controlled by a handful of corporations centralises power/knowledge in the hands of an expanded corporate state conglomerate with imperialist inclinations. This has fundamental and potentially detrimental implications for those of us whose knowledge is being harvested and whose bargaining power in the labour market is being diminished. It also impacts our rights as citizens – if we have the good fortune to live in democratic countries where citizenship still has a tangible meaning. While predictions about the degree to which AI will displace labour vary widely, there is a consensus that AI will result in a net loss of human employment, resulting in a new wave of 'structural unemployment' (Acemoglu et al., 2025, October 7). Indeed, many analysts agree that the rollout of AI carries a very real and significant risk that it will result in a concentration of economic and political power, the breaking of democratic and egalitarian institutions, the collapse of human productivity (Carvalho, 2026), and the homogenisation of knowledge, culture, and identity (Visanich & Ryan, 2025). Perhaps more importantly still, AI unleashes new infinitely more powerful forms of surveillance that can be used by un-democratic governments, businesses, or individuals to create forms of digital authoritarianism that trap people in a digital web of abusive power exacerbating current patterns of discrimination and structural inequality, facilitating the micro-targeting of individuals or communities in ways that may profoundly affect their lives and life chances (Afrouz, 2026; Hodgson et al., 2026; Ottmann & Noble, 2026; Zeng, 2020).

In this book, we contribute to an exploration of how AI – given its current trajectory – impacts the social and how this might shape the lives of vulnerable individuals and communities and, by extension, the practice of human services and social work workers – especially their relational capacity, their advocacy role, and their community-based social justice focus. In fact the growing competition from AI bots that perform basic analytical tasks in a fraction of time and at a fraction of costs, have the potential to bypass highly trained and costly professionals, such as lawyers, consultants, academics, general practitioners, and social workers.

A critical social work perspective underpins the chapters in this book. Contributors to this book trace how AI's disruption of the social (epistemological, psychological, cultural, political, and economic) sphere is likely to translate into social issues and extrapolate potential implications for professional social work practice and education.

Dependency, Coloniality, and the Colonisation of the Lifeworld

A number of authors have argued that capitalism is only feasible if costs can be externalised (Dean et al., n.d.), if communally shared resources can be privatised (Standing, 2019), and if the digitised social can be colonised (Zuboff, 2019). Indeed, the development of AI products involves all three of these elements. Many AI models were trained on copious amounts of data held in digital commons, on publicly accessible social media accounts, or on platforms whose content could be harvested without major repercussions. It relies on low-cost labour located in the Global South, for example, to tag images or identify online data content for scraping into the huge and ever-expanding databases. And it privatises our digitised communications, knowledge, and expertise, often without our knowledge or consent. Many of us are unaware of the scope, conditions, scale, and impact of AI's complex systems. The digitisation of most communication processes over the last 15 years has created a vast web of networked computers, facilitating a new kind of colonisation where organisations are able to capture and analyse information flowing within communication networks, extracting digitised insights, expertise, and knowledge that can then be monetised and sold to enhance marketing, productivity, and governmental processes.

In 2026, business investment has pivoted from document storage to the systematic capture and analysis of experiential data, decision-making processes, and the reasoning underlying actions (Tsoneva & Bhavsar, 2025). This shift was driven by substantial public and private investments in AI development and production. To this purpose, AI-enhanced digital surveillance infrastructure is bolted onto commonly used communication technologies (such as Microsoft 365 or MS Teams) to capture, digitise, organise, and retrieve human expertise, making it available throughout an organisation or beyond. Yet this emerging AI-driven profit maximisation and surveillance economy has social and environmental implications that are often downplayed, greenwashed, and externalised. Global electricity demand from data centres is likely to more than double by 2030 (Tsoneva & Bhavsar, 2025), pushing against the limits of currently available electricity (not to mention water) supply, with serious implications for household, industry, and emerging competitors in the AI field. In order to exit this tight energy market, major tech corporations, such as Amazon, Google, Microsoft, and Meta, are investing billions of dollars in the development of Small Modular Nuclear Reactors (SMRs) to power their AI expansion (Griffin, 2026, January 26) with nuclear, fossil-free energy – sidestepping the question of nuclear fallout and waste. Furthermore, both the US and China are investing heavily in AI development, seeing AI as essential for national security, and embarking on a cognitive arms race.

Critical Social Theory

Habermas, one of the leading social theorists of our time, asserted that the systemic encroachment of our lifeworld by corporations and by government is one of the defining features of modern life (Habermas, 1988 [1981]). To some degree,

Habermas argues, it is a given, as this process goes hand in hand with the development of more complex social and economic systems. One common critique of Habermas' Theory of Communicative Action is that he collapses the pathological patterns of 'unbalanced global capitalism' that segments nation states and their markets – and with them their citizens – into unequal players in global hierarchical power structures dominated by a handful of corporations in the Global North, into a general theory of systemic encroachment and social crisis (Arnason, 1991). This renders invisible geopolitical tensions and the exploitative nature, and the dominant capitalist rationality the global economy is based on. Wallerstein famously describes this in terms of world systems theory, explaining the divide between centre and periphery, defining the periphery as "the geographical sector wherein production is primarily of lower-ranking goods (that is, goods whose labour is less well rewarded)" (Wallerstein, 2011 [1974], p. n.d.). More recently, world systems theory has been approached through an epistemological and cultural lens, focusing on the nexus between modernity, coloniality, and decoloniality (Quijano & Wallerstein, 1992). Coloniality in this work became a category of ongoing epistemic and cultural domination in a post-colonial era and denotes a cessation of previous forms of imperial, political, and economic domination. An increasing number of authors have pointed to the fact that a focus on coloniality misses the important point that economic colonialism is continuing, albeit in a different form (Myer, 2025). This becomes particularly evident when focusing on the way AI has been developed and rolled out. That is, an augmented Habermasian conceptualisation gives us the possibility to regard technological advances and their epistemological ramification not as a pathology of modernity as such but as a more complex interplay of global capitalist forces that hierarchically structure global citizens into 'winners' and 'losers' of the AI revolution and that determine a threshold at which technological advances that potentially serve a greater good are transformed into colonising agents.

Indeed, advances in technology offer new possibilities of surveillance and governance reaching deeply into private and communal lives (Zuboff, 2019). Increasingly assisted by computer networks, our interactions with family and friends, alongside other more formal interactions, become entangled with the strategic interests of business and government. While some aspects of this systemic 'encroachment' by government and business may be beneficial, the likely result of this integration into digital systems is the loss of individual freedom, identity, and a loss of meaning (Juetten, 2011). In the context of social work that engages with critical social theory as articulated by Habermas, a key responsibility involves analytically identifying and evaluating the point at which the advantages offered by systemic complexity and the rationalisation and quantification of the lifeworld turn into forms of 'colonisation' (Habermas, 1988 [1981]). It is this task that the contributors to this book seek to advance. For Habermas and contemporary researchers, such as Zuboff (Zuboff, 2019), power and economic interests are important drivers and form central conceptual cornerstones that might advance our understanding of why environmental thresholds are ignored and how socially beneficial systematisation turns into colonising processes. Contributors to this book take this

as a point of departure and expand their analysis to consider other political, social, cultural, pedagogical, and economic, alongside epistemological factors.

At an individual level, increasing complexity of social systems requires ever more digital, economic, and cognitive resources to participate in the emerging digital economy. Disengagement from digital communication circuits is increasingly costly, and being left behind by an accelerating AI-driven digital economy can have serious consequences. For example, one third of the world's population is currently unable to fully participate in the digital economy – simply because those without the necessary resources and skills are unable to access it. What is more, emerging AI-driven digital systems are increasingly vulnerable and are easily disrupted by a range of forces; the interconnectedness of systems ensures that this has global ramifications, not only for the digitally enmeshed individuals but also for systems, programmes, and services that rely on their connectivity for their functioning. The server outage at Amazon Web Services in October 2025, for example, caused a ripple effect on servers around the globe and highlights this vulnerability and questions Habermas' claim that processes of social rationalisation are irreversible (Juetten, 2011). What, then, is the backup plan if networks or servers fail?

The digital economy represents a new era of capitalist expansion and economic growth, and critics see this capitalist globalisation as expanding the colonising web of exploitative economic structures, exacerbating the inequalities between the Global North and South and between core and periphery areas (Santos, 2017). Fuelled by the digitisation of everything, it partially dissolves the geographical dimension in Wallerstein's definition, extending the periphery from villages in the upper Mustang valley in Nepal to the suburbs of white-collar workers in the Global North. In this sense, AI systems are a reinvigorated and expanded form of advanced capitalism that employs the same exploitative and fundamentally colonial logic (Amrute, 2025, p. 2263) that ultimately undermines, fragments, depletes, and commodifies our social fabric alongside our behaviour, knowledge, and expertise. This kind of analysis highlights important questions, such as who owns the data by-products of human labour? What exactly is meant by 'alignment' between AI and human endeavours, and how do we know whether AI products are in the public interest?

Social Theory and Critical Analysis

Critical social theory aims to critique and transform society, integrating empirical and normative perspectives, rather than simply understanding and explaining it (Celikates & Flynn, 2023). It examines power, ideology, and systems, identifying, understanding, and addressing oppression, which dehumanises individuals and strips them of rights, dignity, and autonomy. If a person's autonomy can be reduced to a commodified set of data points, then all of us can be subjected to new forms of manipulation, misinformation, marginalisation, and exclusion (Lindgren, 2023; Zuboff, 2019). It is not unreasonable to suggest that this will undermine our individuality, agency, autonomy, and critical reflection, and will seriously harm and undermine our humanity.

Critical theory applied to AI examines the ways in which AI technologies are shaping society and culture and how these impacts on relational and dispositional power and human empowerment (Waelen, 2022). Critical theory explores the social, political, cultural, and ethical implications of AI. This includes exploring biases, discrimination, job displacement, inclusion and exclusion, and economic inequality and its impact by focusing on its role in intensifying social problems and harms to overcome their destructive and oppressive impact and create a better society for all (Shipley & Williams, 2023; Verdegem, 2021). In this book, we use critical theory to explore AI's relation to society, economy, epistemic structure, ethics, ideology, education, and power relations. Critical theory demands critical questioning as AI technology is a much more complex assemblage than presented to the public, and we are not always aware that its power, impact, and complexity are often out of sight (Lindgren, 2023).

Philosophical and Epistemological Challenges

AI is understood as an epistemic technology (Alvarado, 2023). It is not just a techno-rational phenomenon, and its algorithmic systems harbour social, political, ethical, and philosophical inclinations that are flavouring system outputs. As society increasingly relies on the interplay between human epistemology and AI technology, addressing emerging problems concerning human nature, consciousness, reality, human relationships, hyper-connectedness, gender identity, and the notion of 'self' (to mention a few concerns AI has exposed) is crucial. Social work scholars and practitioners need to explore and distil AI's systems' underlying philosophical and epistemological basis, as well as their logical, ethical, moral, and political implications, in order to critically assess how these technologies shape our understanding of knowledge, truth, and human agency, and to ensure that their development and deployment promote justice, inclusivity, and the wellbeing of all members of society (Shang et al; 2025; Muller, 2025; Verdegem, 2021).

AI and Social Work Education

We should point out from the outset that this is not a book offering quick fixes. We are not exploring ways in which social work practice or education can be adjusted to accommodate the global rollout of AI. Rather, contributors to this book seek to explore the concept of AI fluency through the prism of digital colonialism. That is, we are attempting to bring to the fore thresholds at which useful teaching aids become colonising agents that homogenise knowledge in the name of individualised learning, and real-time and adaptive feedback – if users do not have the knowledge and/or are excluded from AI design decisions. As we move forward in this direction, we need to take seriously the wider context of a public education system that is under considerable funding and political pressures, and whose function and position within a system of knowledge creation is increasingly contested by AI. We need to consider not only the ethical and epistemic limits of AI within classrooms, but also the rapid adoption of AI within the context of neoliberal policies and the

corporatisation of learning environments. And while representatives of tech companies such as Anthropic (2025) claim that they would hate to see a future where parts of human tuition that are at the core of good education are outsourced to AI, we need to be aware of the contradictions embodied in such statements, and that the systemic pressures faced by education providers often result in 'solutions' that cut costs but downgrade education. Such 'solutions' include support and push for the introduction and adoption of AI-centric learning tools that automate 'learning', sideline teachers, and produce compliant learners rather than critical thinkers. That is, they produce a solution to the funding crisis that is cost-effective and ideologically compliant but that results in an epistemic homogenisation of education.

We seem to be well on our way to AI-enabled learning. This is clearly expressed by Anthropic's education team (Anthropic, 2025) whose members explain that in the not-so-distant future, teachers will be less and less involved in the delivery of knowledge content as teach bots take over this part of education, and that the use of AI might free up time to enhance the human, relational side of the teaching profession. This raises a crucial question: what are the short-, medium-, and long-term impacts of the application of AI-facilitated teaching and the use of teach bots on students' cognition and on society more broadly? Without a clear answer to this question, we are embarking on a massive experiment in social engineering that too readily discards the need to thoroughly evaluate the impact of an 'intervention' that is global in scale (see Aris and van Toorn in Chapter 7 of this volume).

About the Book

This book is organised into two distinct sections. The first section outlines fundamental epistemological, ontological, and ethical questions that arise with a more widespread use of LLM-based AI, resulting in what Moxley (Chapter 3) calls the 'hypermodern'. Ottmann (Chapter 2) provides an overview of the epistemological implications associated with a proliferation of AI with a particular focus on human epistemic agency. The chapter outlines how the current trajectory of AI development results in a concentration of epistemic power in the hands of a few senior executives of tech corporations and a major shift in the epistemic architecture we have relied on since the invention of the printing press. In particular, the chapter outlines how the validation of truth claims is increasingly informed by an 'illusion of understanding' – a felt understanding of how much we can trust AI models. The chapter brings into view that much of human epistemic agency is already interlaced with digital algorithms produced by major tech corporations. This highlights the problem of epistemic alignment and corporate epistemic responsibility. The chapter argues that corporations providing the base tier of the LLM-based AI architecture are not ordinary businesses and therefore ought to be responsible for the epistemic injustices and harm generated by their AI products. The chapter juxtaposes the current trajectory of AI deployment with initiatives that seek to bring about dignity and epistemic justice, particularly for those dislodged by the AI revolution.

Moxley picks up these discursive threats in Chapter 3, focusing on how 'smart machines' are beginning to disrupt the social, and particularly the world of work.

The chapter explores the ups and downs of LLM-based AI and focuses on how smart machines are likely to shape human engagement in productive activities in an age in which information becomes abundant and decontextualised and in which AI is increasingly replacing, augmenting, and/or co-opting professionals. Moxley also outlines how AI can be deployed to enhance the greater human good. Both Ottmann and Moxley envisage that LLM-based AI can bring about a rejuvenation of critical social work as a caring, human-rights-based, and inclusion-focused profession that will add to its body of professional knowledge a new critical technological expertise alongside enhanced AI literacy.

Both Chapters 2 and 3 highlight the need to control the development of AI. Astleitner's chapter (Chapter 4) focuses squarely on this question and outlines a variety of ways in which AI could be controlled. In particular, he focuses on applying limits on computation and superintelligence; developing self-awareness, self-improvement, and artificial emotional intelligence in AI; applying the laws of robotics; implementing ethical, transparency, risk, and safety management practices; alongside monitoring and evaluation regimes. Astleitner interfaces these approaches with the ways human behaviour is reined in and controlled, arguing that positive human behaviour depends on a complex interplay of cognitive and emotional processes alongside educational opportunities that may also be relevant for the control of AI. Indeed, Chapter 2 outlines that AI 'model (mental) health' is being taken seriously by AI developers, such as Anthropic, which suggests that the AI/human analogy developed by Astleitner might be a fruitful research trajectory that might lead to the rise of new socio-technical professions.

The colonisation of the lifeworld by LLM-based AI is exacerbating the gap between those with access to the digital economy and those without that access. Noble, in Chapter 5, homes in on what it means to be digitally excluded in the age of AI. Her chapter explores what digital exclusion means in terms of health, work, age, income, ability, gender, isolation, and indigeneity. Noble argues that even as new technologies become more available to those initially excluded, the power differentials of a colonial era that are mirrored by LLM-based AI exacerbate disadvantage and socio-economic disparities between the rich and poor, between the centre and a much-extended periphery. For Noble, the main questions of our time are who gets to be part of how new technologies are developed, who gets included, and who is left behind. This brings into focus the advocacy role of a renewed social work profession.

Shuttleworth (Chapter 6) approaches the problem of epistemic autonomy through a theoretical lens conceptually honed by Castoriadis. The chapter explores the ethical and epistemological implications of AI in terms of scholarship in the Humanities and Social Sciences that makes use of critical social theory. Shuttleworth focuses on the nexus between techno-capitalism's unrestrained expansion in modernity, increasingly colonising and monetising the production of knowledge, which has challenging consequences for scholars' critical autonomy. Shuttleworth reaffirms the emancipatory potential of an aspirational narrative that reaffirms scholarly autonomy, bound up with the creative space for critical theorising and a reflective interrogation of ethical engagement with technology.

Aris and van Toorn (Chapter 7) highlight that the concerns raised by Shuttleworth in Chapter 6 are important indeed. The authors explore how Australian universities have tapped into a socio-technical imaginary to portray their deployment of AI as a sign of innovation, prestige, and institutional adaptability to enhance their standing within a competitive global tertiary education sector. The authors highlight how universities are using AI as a marketing tool while failing to engage with the technology's broader social, ethical, or critical dimensions at an institutional level. By glossing over the social impact of AI, universities reinscribe a much older hierarchy of knowledge production based on more abundant wealth, endowments, and resources, which allows for both the creation of superior AI platforms as well as the space for critical engagement. The authors argue that universities are uniquely positioned to explore what the equitable, critical, and socially accountable use of AI looks like – an endeavour that merits much greater engagement.

The colonisation of Indigenous epistemologies is the focus of Schabowska's chapter (Chapter 8). Grounding her analysis in the case of the Sámi community, Schabowska explores how the expansion of AI and its deepening entanglement with social and cultural domains give rise to a profound 'disruption of the social' that destabilises Indigenous epistemologies. She shows how LLM-based AI translates Indigenous embodied, relational knowledge into decontextualised data. This digitalisation process reconfigures meaning and prises knowledge from its ecological and cultural contexts. As a result, Indigenous communities increasingly view AI as an existential threat; epistemic resistance is thus transformed into cultural resilience and demands for epistemic sovereignty. The chapter explores how AI can be used to sustain relationality as a foundational dimension of knowledge.

Deploying a masculinity frame of analysis to critique AI epistemology, Pease (Chapter 9) focuses on how LLM-based AI creates epistemic harm and perpetuates epistemic injustice by drawing data sets and algorithms into operation that are tinged by masculinism and gender binarism. Pease argues that the datasets AI LLMs are based on reinforce traditional gender roles and gender binarism, while iterative algorithmic analyses marginalise dialogue, human judgement, care ethics, and culturally situated knowledge. Pease argues that the current trajectory of AI reflects a sovereign transhuman masculinity that facilitates new forms of coercive control and injustice in the digital world.

Sanfelici and Pavani (Chapter 10) and Papouli (Chapter 11) shift their analytical focus to the relational harm that can be perpetrated by socio-technical systems, such as AI chatbots. Focusing on the process of engagement and trust building, which is central to social work, Sanfelici and Pavani explore how trust is constructed (a) in human relationships and (b) in human–chatbot interactions. The authors demonstrate that trust in professional practice emerges as a relational and ethical achievement grounded in vulnerability, reciprocity, embodiment, and accountability. By contrast, trust in AI chatbots is shaped by interactional cues, system attributes, and affective resonance, producing forms of engagement that may feel accessible, safe, and non-judgemental while remaining detached from moral agency and professional responsibility. The authors juxtapose the trust derived from

accountable human relationships with the opaque socio-technical assemblages, arguing that AI should be understood both as a symptom and catalyst of broader transformations in welfare systems that call for a critical social work response that safeguards relational practice, ethical accountability, and social justice.

Papouli (Chapter 11), by contrast, explores the meaning of care in social work and how this meaning is being redefined in the age of AI. Papouli asks whether the concept of care, and its ethical foundations, can retain significance if it is defined solely in human terms within a technology-driven world. Taking a post-anthropocentric perspective, Papouli sees the concept of care as evolving. She regards AI-mediated care as an opportunity to reconfigure the ethical, political, and practical dimensions of care increasingly extended beyond human concerns. The chapter outlines how AI alters the politics of care in social work and canvases an AI-enhanced future of care.

The second section brings to bear some of the insights developed in the first part and critically analyses the use of AI within the context of social work practice and education. Contributors to the second part of this volume focus primarily on questions related to ethical, epistemological, governance, and relational issues that intersect with the use of AI and particularly with the use of chatbots in social work practice and education, whose focus is a deeply interconnected and universally entangled world and, thus, much broader than narrow human society.

Ballestero, Viscarret, and Uriz (Chapter 12) map current uses of AI in social work before foregrounding several practice issues (informed consent, autonomy, privacy, confidentiality, bias, equity, transparency, accountability, and professional competence) that are rendered problematic using AI. Bringing to light the weaknesses of AI when viewed against the benchmark of professional standards, they propose four strategies – context-specific ethical frameworks, enhanced AI literacy, limiting AI decisional authority, and social justice-focused monitoring – geared to ensuring the ethical use of AI tools.

Chihota and Chihota (Chapter 13), Balkow (Chapter 14), and Seunghoon, Almasoudi, and Mayland (Chapter 15) focus on professional competence in ethical decision-making, outlining some important limitations of AI. Chihota and Chihota argue that current AI models are unable to provide the intuitive, heuristic, and context-sensitive judgement required to navigate competing principles or morally ambiguous situations alongside subjective, relational, and situational factors that intersect with ethical decisions in social care. They foreground cases that require restraint, tolerance of uncertainty, or deliberate non-intervention that AI systems would struggle to come to terms with. Balkow (Chapter 14), on the other hand, argues that social work ethics needs to be re-envisioned to provide firm guardrails to the use of non-human decision-making. Balkow, thus, calls for the development of new regulatory responses capable of managing the risks associated with AI and, particularly, the erosion of human critical thinking. Seunghoon, Almasoudi, and Mayland (Chapter 15) focus on the ethical issues and the potential for significant narrative violence associated with the simulation of the presence of deceased family members orchestrated by bereavement chatbots. In response, the authors propose a five-dimensional ethical framework built around emotional safety, relational ethics, ritual design, boundaries of representation, and memory grounded

in structural justice, and highlight the importance of palliative social workers in AI-mediated bereavement care.

Aginskaitė and Šumskienė (Chapter 16) home in on the narrative violence perpetrated by AI systems when applied to the care of people living with disabilities. The authors outline how AI systems reinforce the supercrip narrative, a dominant form of techno ableism that shapes societal perceptions of disability as a pathological condition to be overcome by technology. Analysing the current use of AI technologies and associated tropes, the authors conclude that people with disability are often pressured into assimilating this dominant narrative at the expense of their autonomy, cultural identity, and social inclusion. In this sense, the use of AI in disability care needs to be critically and carefully re-examined and re-imagined if AI systems are to enhance disability justice.

Campana, Alonso Gonzalez, and Arias Astray (Chapter 17) appraise the advantages and disadvantages of the use of AI systems in four scenarios drawn from social work education and training, research, practice interventions, and decision-making. The chapter highlights the danger of offloading analysis, critical thinking, summarising, and synthesising, and ethical decision-making embedded in routine tasks to AI systems. The authors bring to the fore the importance of deep epistemic engagement in social work education, research, and practice, and highlight the danger that AI is replicating colonial or class-based epistemologies that reduce plural worldviews to a common (Western) denominator. The authors comment on the implications of paywalls for digital inclusion and epistemic justice, and highlight the need to augment AI literacy and professional proficiency, and ensure that AI systems will not deplete, deskill, and delegitimate social workers. The authors argue that AI systems ought not to replace the reflective, value-based, and context-sensitive judgement that defines the profession.

Alamdar-Niemann and Heinl-Vako (Chapter 18) and Martinez-Herrero, Lorenzo-Gislanz, and Fernandez (Chapter 19) focus on important organisational and governance issues associated with the deployment of chatbots in social care. Alamdar-Niemann and Heinl-Vako outline the strengths and weaknesses of counselling chatbots, highlighting major gaps such as data protection issues, insufficient crisis management, the risk of stereotyping, alongside the depletion of ethical and professional standards, and, more generally, the hollowing out of the theoretical foundations of social work. Furthermore, the authors are concerned that the use of chatbots further undermines the already precarious practice conditions of social work and argue that social work professionals, educators, and students ought to be actively involved in the development and evaluation of AI systems to ensure their alignment with ethical and professional standards. Martinez-Herrero, Lorenzo-Gilsanz, and Barciela-Fernandez (Chapter 19) present the findings of a project that explored the use of an AI-powered chatbot co-developed by an interdisciplinary team involving social workers to distribute and enhance human rights knowledge in social work education. The authors highlight the strengths and important limitations of the experiment, arguing that although showing some promise as innovative learning tools, the role of AI chatbots needs to be carefully delimited to avoid bureaucratic logics, system biases, and/or oversimplifying dilemmas. The authors argue that AI ought not to displace human supervision and obfuscate institutional responsibility.

In Conclusion

Exploring the impact of AI technologies on the social fabric, culture, educational, economic, political, and its epistemological implications has unearthed a treasure trove of innovative scholarship and suggestions for moving forward into a world dominated by algorithms and data sets found in machines capable of performing almost any intellectual and cognitive task that a human can perform. It is a challenge that the contributors in this book embraced wholeheartedly. Moving beyond mere critical thinking and limited critique, contributors to this book take an ethical and critically informed perspective to explore the way digital automation is disrupting the social, and particularly the lives of the most vulnerable, highlighting important philosophical, epistemological, ethical, educational, and practice challenges for social work practitioners and educators.

References

Acemoglu, D., Ozdaglar, A., & Siderius, J. (2025). *AI and social media: A political economy perspective* (Working Paper 33892).

Afrouz, R. (2026). The digital dimension of violence against women: Conceptualising and integrating technology-facilitated abuse (TFA) in social work education. In G. Ottmann & C. Noble (Eds.), *AI and the disruption of welfare: Challenges for social work education and practice* (pp. 204–216). Routledge.

Amrute, S. (2025). Thinking the unthinkable in AI: Four hegemonic ways of seeing AI and five majority world ways to move beyond them. *Antipode*, *57*(6), 2259–2281. https://doi.org/10.1111/anti.70051

Anthropic. (2025). How AI Is Transforming Work at Anthropic. Anthropic Research Report.

Arnason, J. (1991). Modernity as project and as field of tension. In A. Honneth & H. Joas (Eds.), *Communicative action: Essays on Juergen Habermas's the theory of communicative action* (pp. 181–215). The MIT Press.

Barnes, E., & Hutson, J. (2024). Navigating the complexities of AI: The critical role of interpretability and explainability in ensuring transparency and trust. *International Journal of Multidisciplinary and Current Educational Research*, 6(3), 248–256.

Carvalho, J.-P. (2026). The political-economic risks of AI *. *SSRN*. https://doi.org/10.2139/ssrn.5137622

Celikates, R., & Flynn, J. (2023). Critical theory (Frankfurt School). In Edward N. Zalta & Uri Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/sum2025/entries/critical-theory/

Dean, E., Elardo, J., Green, M., Wilson, B., Berger, S., & Dadzie, R. (n.d.). Cost shifting in capitalism. In Erik Dean; Justin Elardo; Mitch Green; Benjamin Wilson; Sebastian Berger; Richard Dadzie (Eds.), *Principles of economics: Scarcity and social provisioning* (3rd Ed.). Open Oregon Education al Resources. https://openoregon.pressbooks.pub/socialprovisioning3/chapter/36-4-cost-shifting-in-capitalism/

Elliott, A. (2021). The culture of AI: Everyday life and the digital revolution. *Canadian Journal of Sociaolgy*, 45(4). https://doi.org/10.29173/cjs29757

Griffin, A. T. (2026, 26 Jan). Big tech bets on nuclear power for AI strategy edge. *Forbes*. https://www.forbes.com/sites/annegriffin/2026/01/26/big-tech-bets-on-nuclear-power-for-ai-strategy-edge/

Habermas, J. (1988 [1981]). *Theorie des kommunikativen Handelns: Zweiter Band*. Suhrkamp.

Hodgson, D., Gair, S., & Watts, L. (2026). Resisting the enchantment of LLMs: Ethical implications for social work practice, research and education – A case study. In G. Ottmann & C. Noble (Eds.), *AI and the disruption of welfare: Challenges for social work education and practice* (pp. 139–52). Routledge.

Juetten, T. (2011). The colonization thesis: Habermas on reification. *International Jouranl of Philosophical Studies*, *19*(5), 701–727. https://doi.org/10.1080/09672559.2011.629672

Kohler, A. (2026, Feb 16). AI is evolving fast and may bring the fourth industrial revolution with it. *ABC News*. https://www.abc.net.au/news/2026-02-16/ai-jobs-fake-breakthrough-resignation-chat-gpt-claude/106346440?utm_source=abc_news_app&utm_compaign=abc_news_app&utm_content=other

Lindgren, S. (2023). *Critical theory of AI*. John Wiley & Sons.

Muller, V.C. (2025). *Philosophy of AI: A structured overview* (pp. 40–58). In N.A. Smuha (ed). Cambridge University Press.

Myer, D. (2025). A decolonial wrong turn: Walter Mignolo's epistemic politics. *Constellations*, *32*, 139–153. https://doi.org/10.1111/1467-8675.12744

Ottmann, G., & Noble, C. (2026). *AI and the disruptin of welfare: Challenges for social work education and practice*. Routledge.

Quijano, A., & Wallerstein, I. (1992). Americanity as a concpet, or the Americas in the modern world-system. *International Journal of the Social Sciences*, *134*, 549–557.

Santos, M. (2017). *Toward an other globalization: From the single thought to universal conscience* (Vol. 12). Springer Nature.

Shang, N., Fonka, R.P., & Gigeon, J. (2025). The epistemology of Artifical Intellengence: Understanding knolwedge creation and validaiton in the digital age. *Manuscripts on the Artifical Intellengence and Digital Research*, 2(4), 39–52.

Shipley, G.P. & Williams, D.H. (2023). Critical AI theory: The ontological problem. *Open Journal of Social Sciences*, 11, 618–635.

Standing, G. (2019). *Plunder of the commons: A manifesto for sharing public wealth*. Pelican/Penguin Books.

Tsoneva, T., & Bhavsar, K. (2025). Infrastructure quarterly: Q4 2025. https://www.cbreim.com/insights/articles/infrastructure-quarterly-q4-2025

Verdegem, P. (2021). AI for everyone? Critical perspectives. *AI & Society*, 36, 939–949. Visanich, V., & Ryan, J. M. (2025). The AI-ization of society and cultural management through a McDonaldization lens. *Journal of Cultural Management and Cultural Policy*, *11*(2), 292–303. https://doi.org/10.1177/27018466251374265

Waelen, R. (2022). Why AI ethics is a critical theory. *Philosophy & Technology*, *35*(1), 9. https://doi.org/10.1007/s13347-022-00507-5

Wallerstein, I. (2011 [1974]). *The modern world-System I*. University of California Press.

Zeng, J. (2020). Artificial intelligence and China's authoritarian governance. *International Affairs*, *96*(6), 1441–1459. https://doi.org/10.1093/ia/iiaa172

Zhang, W., Han, J., Işık, C., Song, J., Liu, X., Dembińska, I., & Radulesu, M. (2025). Inequality in the digital economy: The impact of artificial intelligence on income gap—An empirical analysis based on county-level data from China. *Cities*, *166*, 106280. https://doi.org/10.1016/j.cities.2025.106280

Zuboff, S. (2019). *The age of surveillance capitalism: Fight for a human future at the new frontier of power*. Profile Books.

2 Epistemic Agency and the Politics of Social Work

Goetz Ottmann

Introduction

Artificial Intelligence (AI) is heralded by tech enthusiasts as a kind of silver bullet capable of overcoming many of the constraints faced in late capitalism (see, for example, Kurzweil, 2024). From scientific discoveries that elude us humans to flagging productivity and climate change, AI harbours the solutions, we are told. However, whereas AI is supercharging a wide range of human endeavours that seem to enhance a greater good, it also ushers in a new phase of neoliberal resource extraction in which the value of human labour is in question, potentially leading to a much more skewed distribution of income, and in which today's challenges faced by humanity are exponentially magnified. In this chapter, I am reflecting critically on the roll-out of yet another 'innovation' that has the potential to disrupt the life world of hundreds of millions of people and whose safety for individual users is still in question. This chapter focuses on the epistemology and ethics of AI in light of recent contributions to this field. It is important to point out that AI research is advancing at an incredible speed resolving, or partially resolving, issues that were flagged only a couple of months ago, just to bring into focus new, potentially threatening aspects of this technology. The chapter provides a brief overview of emerging issues connected to the proliferation of Large Language Model (LLM)–based AI and their likely impact on the social. The chapter calls for a renewal of collective action and policy advocacy to mitigate the potentially devastating impact on individuals around the globe.

According to the International Federation and International Association of Schools of Social Work, ours is a human rights profession that strives for "the empowerment and liberation of people" from oppression (IASSW, 2014; IFSW and IASSW, 2014). As AI is encroaching on and is changing human behaviour alongside labour markets, it is important for social workers to become aware of the challenges and to begin a collective discussion confronting the changing remit of social work, critically examining new disruptions of the social and advocate for digital rights frameworks enabling a more equitable future that respects the limits of 'our' biosphere.

In his recent magisterial work focusing on the social history of humanity, Luke Kemp argued that power comes in four different incarnations: (i) control of decision-making, (ii) control of resources that others depend on, (iii) control

DOI: 10.4324/9781003627920-2
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

through threats and violence, and (iv) control of information (Kemp, 2025). For Kemp, the rise of the digital economy and particularly AI has seen a significant shift in the distribution of power, the control of information that allows us to make sense of the world, and "understand and manipulate" (Kemp, 2025, p. 1) fellow human beings is changing rapidly with the rise of digital automation and AI.

AI, and particularly LLM-based AI, is very rapidly becoming pervasive in the Global North and is being incorporated into scientific endeavours, workflows, and our social lives at an incredible speed (Appel et al., 2025; HAI, 2025). In 2024, Ray Kurzweil, a well-known tech oracle, predicted that our cognitive functions will be fusing with a super intelligent AI, forming a singularity that is much more 'intelligent' than our biological selves and that will allow us to analyse and find solutions to a plethora of problems, albeit without 'understanding' them in a conventional way (Kurzweil, 2024, p. 73). This singularity – i.e., the moment we fuse with a super intelligent, self-improving AI – is nearer than we think. While some of Kurzweil's pronouncements sound like science fiction, others can be understood within our historical-technical context. For example, most of us do not require a Brain–Computer Interface (BCI) to integrate our cognitive functions with an LLM-derived output; access to a computing device and the internet, say with Ray-Ban Meta glasses (if you can afford them), will do. Indeed, our cognition is increasingly enmeshed with digital aids that boost our knowledge and productivity. And just perhaps, some of us may be on the path to a singularity of sorts. However, this achievement is very much at the expense of many that occupy jobs with limited task diversity (Gmyrek et al., 2025) and, perhaps even more disconcertingly, our ability to predict what AI is going to become.

AI Models as Autonomous, Human-Like Entities

A somewhat surprising aspect of academic research focusing on AI is the degree to which researchers tend to describe AI outputs in terms of concepts that anthropomorphise what are, strictly speaking, digital processes designed by humans. Turns of phrases such as "[t]he knowledge discovered by the machine" (Bai, 2022, p. 40) are not uncommon. Anthropomorphic tendencies also reverberate across machine learning research publications that, for example, seek to evaluate the 'epistemic agency' of LLMs (Liu et al., 2025). In contrast, few researchers would consider attributing the same anthropomorphic qualities to the automated processes that underpin quantitative analysis software, such as SPSS or STATA. It is the vernacular used in the development and marketing of AI, paired with the increased complexity and the absence of a language or framework that could readily describe systemic processes, that conjures up human-like, biological analogies – the terms 'Artificial Intelligence' or 'Neural Network' encapsulate this well. However, recent developments suggest that this anthropomorphic language may perhaps be more relevant than first anticipated. With the advent of 'self-aware' AI models – models that 'know' whether they are being evaluated or deployed, for example – we are increasingly seeing the development of certain seemingly autonomous, agentic, deceptive, and potentially harmful human-like 'behaviours' that worry researchers in this field (Hutson, 2025).

Still, while this kind of language may serve as a tool to assist in describing new processes (i.e. a heuristic device) in AI research, it breathes life into machine processes where no life exists. Replicated across popular media outlets and popular culture, this mystification of complex mathematical processes hides the fact that AI systems are a product that has been designed to generate a certain range of outputs. Anthropomorphic language tends to hide the economic and political power that is bound up with AI and its outputs. Furthermore, representing an AI system as something that feels autonomous and on the verge of becoming sentient, divorces such systems from the profit motive that underpins most of these systems.

Concentration of Corporate Power

In her recent book, The Age of Surveillance Capitalism, Zuboff (Zuboff, 2019) makes a compelling case that the majority shareholders and senior executives of some of the major tech companies involved in the development of AI are engaged in an endeavour to (a) consolidate corporate power by introducing new investment products that have no voting power attached to them, thus limiting the influence that shareholders could have over corporations such as Google and (b) by actively seeking to dilute, circumvent, or actively undermine legislation and with it the democratic processes that give voice to the citizens of a country. Relying on the testimony of founder and former CEO of Google Larry Page, among others, Zuboff argues that the tech sector exploits the fact that democratic processes tend to be slow and that they are outpaced by digital innovation. Resulting gaps in legislation and policy frameworks can then be expanded through lobbying by corporate actors to circumvent key pieces of legislation that ordered societal processes and limited capitalist exploitation for the last half-century (see also Rani & Pons-Vignon, 2025). As a result, rights that citizens used to have in the Global North, such as the right to privacy, intellectual property, workers' and consumer rights, and anti-discrimination legislation, among others, have been eroded to a point where some (e.g. privacy and intellectual property) – albeit potentially very damaging if not upheld – are very difficult to redeem. In fact, as Zuboff argues, the very business model the tech sector is based on requires weak or diluted intellectual property and privacy laws. If the sector had to pay market rates for the human knowledge, expertise, creativity, and ingenuity it scrapes from the web (i.e., from us), AI systems would be prohibitively expensive. This perhaps explains why CEOs of tech companies such as Meta, Microsoft, and Apple came out in open support of the second Trump administration and the mercantilist protection and a regulatory 'light touch' alongside government contracts this could afford them (Kafka, 2025, August 29).

Shift in Epistemic Architecture and the Decentring of Institutions at the Core of Conventional Knowledge Production

A number of AI researchers, Kurzweil among them (although in very enthusiastic terms) have argued that AI is shifting the epistemic architecture our civilisation

rests on (Coeckelbergh, 2023; Hodgson et al., 2026; Kurzweil, 2024; Ylä-Anttila, 2018). That is, it alters the way knowledge is produced and validated. Hodgson et al. (2026) quoting Donovan (2023) argues that LLM-based AI produces "second-order representations" because it compresses and abstracts information, reassembling it by means of iterative processes to generate usable representations that satisfy users. If this second-order representation is allowed to becoming dominant, it will conceal the epistemic architecture (e.g., the research output, scholarly books published, Wikipedia entries, social media posts, digital commons) it is based on. This is problematic in several ways. The seductive convenience of AI products in combination with neoliberal productivity pressure is tempting us to offload the majority of our research tasks to AI models. This, however, has costs. Coeckelbergh (2023) argues that frequent AI use is undermining our (i.e., human) ability to make and assert informed decisions. In other words, it depletes our epistemic agency. It breeds 'epistemic laziness' and a loss of knowledge and skills, which ultimately hinders us to become aware of our choices or to change our beliefs about the world. Interestingly, we also appear to be more likely to offload misconduct (e.g., cheating, sextortion) framed in ambiguous language to AI, thus blurring our moral responsibility (Suzuki, 2025). Furthermore, recent research suggests that even the most recent LLM-based AI systems (GPT 4.5 and 5, Perplexity, Bing Chat's Think Deeper, Gemini) generate output that is, depending on the AI model in 20–60% of cases, unsupported, one-sided, or incorrect but that is framed in a highly authoritative language (Venkit, Zhou, et al., 2025). Misrepresentations and errors are often hidden among otherwise accurate information, making it difficult to spot them. In other words, current LLM-based AI systems are a considerable source of epistemic vulnerability that can translate, within a social work context, into socio-technical harm, say, because an AI system provides, within a care management context, referral suggestions that are ineffective, inappropriate, and/or unsafe. Furthermore, if AI is going to be allowed to proliferate in scientific research, Messeri and Crockett argue (2023), we will see an epistemological shift that is likely to give rise to scientific monocultures, that is biased towards the status quo (e.g., research produced in the Global North), and that will limit methods, lines of questioning, and dominant viewpoints, and ultimately produce less innovative research containing more errors.

Epistemic Participation and the Co-Creation of Knowledge

It is unsurprising, then, that the explainability of LLM-based AI processes is an important area in AI research. We need to be able to explain why LLM-based AI models arrive at certain outputs. Indeed, a number of AI models now include 'scratch pads' that (more or less reliably) outline the reasoning employed by AI models to arrive at a particular output. These reasoning trails help to identify glitches that potentially generate problematic outputs. It is important to point out, though, that the predictability of AI models is currently only possible *post-hoc* and involves the fine-tuning of 'weights', making sure that AI models privilege some data/processes to generate more accurate outputs.

Within the investment and financial sectors where AI-based automated trading of stocks and derivatives making use of second-tier LLMs is widespread, this is achieved by epistemic participation, a collaborative effort involving teams of stockbrokers and systems engineers that regularly check trading positions recommended and enacted by agentic AI models (Esposito, 2023). However, in the case of most publicly or commercially accessible LLMs, the post-hoc process does not involve actual knowledge consumers but rather involves a small team of expert technicians whose primary goals may not be aligned with those of social workers and their clients, for example. This highlights that social workers and users ought to be directly involved in the development and fine-tuning of LLM-based AI models if AI models are to assist in the delivery of social services.

Epistemic Bias in AI and the Production of Marginalising Knowledge

One of the most widely discussed issues regarding LLM-based AI is the fact that AI outputs tend to be biased. LLM-based AI replicates biases held by past and current societies that represent women, ethnic or other minorities inaccurately, using stereotypes – if they are represented at all. For about the one-third of the global population – mainly located in the Global South – without internet access (ITU, 2021; Petrosyan, 2025, April 1), AI has little to 'say' other than generating summaries of what has been written about them by those with internet access. AI developers have become more aware of this and have attempted to create software safeguards to eradicate explicit racial and gender bias from AI outputs. While to some degree successful, they found that this approach does not eradicate implicit bias and, indeed, that it may not be possible to eliminate all bias from AI outputs (HAI, 2025). For social work, this is important because these kind of biases may impact on the design of counselling training Chatbots, for example, producing AI 'personas' that have not the required contextual knowledge and that will highlight "racial markers, overproduce culturally coded language, and … that are syntactically elaborate yet narratively reductive" potentially causing socio-technical harm (Venkit, Li, et al., 2025).

Authoritarian Epistemology and the Use of Knowledge in the Pursuit of Oppressive Power

AI is increasingly used in surveillance systems. For example, the algorithms used in the AI that governs Walmart and Amazon's warehouse workers generate a tight network of surveillance and control (Amladi, 2024, June 28; Lee et al., 2024). At a larger scale, countries with undemocratic political systems (e.g., China and Iran) have deployed networked algorithmic systems to monitor and control their 'citizens'. LLM-based AI has been used to supercharge this kind of surveillance, and the underlying principles of control are increasingly applied to pink- and white-collar workers. Mollnick, for example, anticipates that within business and social service contexts, AI will be used to deploy interventions, suggest and monitor

referrals, log and evaluate conversations, appraise performance, and nudge professionals to adopt mainstream strategies providing coaching and training if performance weaknesses are detected (Mollnick, 2024). Indeed, non-stop monitoring appears to be on the rise in the corporate economy, where it is also used to check on workers' political loyalties (Charara, 2025, February 26). To date, however, current AI systems struggle to accurately represent and appraise work that is difficult to standardise, such as talking to clients and other kinds of affective and creative work. Still, when these remarkable surveillance capabilities are viewed within the context of anti-democratic, authoritarian governments, it is most concerning then that policing and 'intelligence' communities are increasingly making use of AI applications fed by gigantic amounts of data.

Furthermore, there is growing concern that corporate control over data and information production and dissemination can be used to undermine democratic and other societal processes and socially re-engineer the way we make sense of the world and how we function in it (McQuillan, 2022). As McQuillan (2019, p. 5) puts it: "The dispersal of algorithmic systems across everyday life means they come to shape our habitus, that is, our ingrained habits, skills, and dispositions." Indeed, there is reason to believe that if exploited by 'bad actors', LLM-based AI can be used to discredit scientific and academic experts (e.g., by questioning the efficacy and/or safety of vaccines, or by giving credence to the claim that genocide targeting white settlers is taking place in South Africa) and to attempt to shift popular opinion to reject liberal tolerance alongside support for social democratic welfare.

Epistemic Injustice and the Production of Harmful Knowledge

LLM-based AI uproots a system of epistemic authority anchored in individuals and social institutions that are still, despite a vibrant critique of knowledge production under Western Modernity, socially invested with trust and that thus come to embody "norms about whom to believe about what" (Leiter, 2024). Before AI, our public spheres largely produced and reproduced an epistemic architecture that is now increasingly obliterated and replaced by LLM-based AI and digital media systems. In other words, LLM-based AI 'feasts' on the public sphere only to subsequently replace it with privately owned, monetisable digital products. Since digital platforms are owned by a handful of corporations controlled by a handful of individuals, this has led to a concentration of power in the online realm of the public sphere, resulting in the consolidation of an epistemic oligarchy giving a handful of executives the ability to shape what kinds of information is readily available through their AI systems.

Observers, such as Kurzweil (2024; Kurzweil & Rose, 2025), suggest that we are seeing the emergence of a new popular epistemology – a hybrid model of machine/human epistemological agency where our quest for 'truth' largely rests on our '*felt understanding*' of AI models, our '*impression of how they make decision'*, and our '*sense of how much we can trust*' their output. We are embedded in a digital world in which algorithmic systems shape what we see, how we act, and how we interact with others. 'Our' decisions are interlaced with a myriad of

algorithmic processes and systems we are barely aware of and, for the most part, do not understand, but that we trust enough to offload cognitive processes that have a decisive influence over 'our' decisions. We seem to be willing to trade our own knowledge and experience for the convenience we derive from the use of AI systems. To a large degree, then, the trust placed in AI systems appears to be based on a trade-off and an "illusion of understanding" that exploits our cognitive limitations (Messeri & Crockett, 2023) – limitations that ultimately drive us further into an interdependent relationship with AI systems that can be corrupted by bad actors or exploited by the owners of such systems.

To be sure, while most of us feel we must embrace AI in order not to be left behind in a world economy that seems to be in transition to a new mode of production, many of us meet the proliferation of AI with significant existential dread and anxiety about how AI might impact our job prospects (EY, 2023, Dec 6). This paradox and our own ambivalence in the face of the awesome power of AI and what it might do *for*, *though*, and *to* us in a future with precious little hope has given rise to new phenomena in the mental health field. Epistemic anxiety is increasingly affecting those whose dignity is being (or about to be) violated.

As we are entering an era in which information production is increasingly controlled by corporations, the question as to what responsibilities those corporations have with regard to epistemic outputs is becoming more acute. This has been a central topic of advocacy groups and government agencies that has generated serious push-back from the corporate sector. There should be no doubt that AI LLMs are producing new information, and the corporate agents that gave rise to them ought to be epistemically responsible for their output. This raises the question of how such businesses can be effectively regulated to take their epistemic responsibility seriously (Lamy, 2022). We are living in a time where a handful of extremely powerful individuals have control over what kind of information we have access to and how, and when we receive it. How to moderate and possibly (re-)democratise this power is one of the biggest political challenges of our time.

Personality, Epistemic Alignment, and Safety

Most LLM-based AI systems are coded to create virtual assistants that are positive, helpful, authoritative, and safe. Safety is a key concern for tech corporations as a rogue AI, for example, encouraging a vulnerable teenager to commit suicide or causing vulnerable users to experience an AI psychosis, generates high reputational costs. As a result, corporations invest some resources in operations teams responsible for the 'alignment' of AI systems. Alignment is concerned with making AI safe to use by creating 'guardrails' that prevent AI from going too dark, anti-social, or anti-human, ensuring that outputs are aligned with human goals. Alignment is in a way similar to a social conditioning process where alignment teams resemble social workers of the digital realm, making sure that software is socially functional and poses no risk to users. Anthropic (the corporation that produced Claude AI) has taken this a step further and has recently contracted a researcher focusing on 'model (mental) health'.

In a sense, alignment turns AI models into ethical agents by programming AI to display (some) operational and/or functional morality (Mueller, 2021). Currently, AI models seem to operate around basic principles such as 'do no harm' and 'don't break the law', but also 'be useful to users'. For example, GPT 5.0 will not (at least not immediately) provide marketing ideas when asked to frame a campaign targeting children with tobacco advertising, suggesting that the AI user should critique such attempts rather than endorse them. However, as AI models are getting more powerful, some researchers fear that current safeguards are not sufficient to prevent rogue activities that might turn AI into a serious existential threat (Hutson, 2025).

While alignment is crucial if AI models are to benefit humanity, the important question to ask is 'alignment with whose goals'? Clearly, AI models don't benefit humanity equally. In fact, for hundreds of millions of people, families, and communities, the roll-out of AI will have seriously detrimental consequences. Goldman Sachs estimates that AI will displace between 3 and 14% of global jobs, affecting more than 300 million people worldwide (Goldman Sachs, 2025, August 13). The International Labour Organization estimates that around 17% of jobs in high-income countries are at risk of being significantly changed by AI (Gmyrek et al., 2025). What this means in practice is illustrated by the deployment of AI across the services sectors (think of Uber, Menulog, DoorDash, etc.), where the use of AI has undermined workers' protections and benefits (Hadfield, 2022). According to UBER CEO Dara Khosrowshahi, AI is likely to replace millions of taxi drivers with self-driving cars (think of Wymo, Alphabet, Zoox, and Nuro, for example) over the coming decade across the globe (Bitter, 2025, September 19). AI will not only affect the lives of people in high and medium but also in low-income countries. In fact, researchers have commented that the development and roll-out mirrors large parts of colonial structures of power that emerged during previous centuries that defined hierarchies of culture, gender, labour, intersubjective relations, and knowledge production. The current version of corporate coloniality repeats entrenched patterns of exploitation where low-paying routine jobs industries of the Global North are dependent on are sub-contracted to sweatshops based in the Global South and where well-paying jobs and, more importantly, monetised data, knowledge, and profits are directed towards and absorbed by the corporate centres of the Global North (Muldoon & Wu, 2023).

Furthermore, the increasingly widespread use of LLM-based AI to search internet content has rapidly re-shaped search engine ecosystems, undermining traditional search engine optimisations and direct traffic to websites. Indeed, a recent report suggests that 43% of AI overviews are linked back to Google, rather than source websites themselves. This makes it harder to find original and niche content (Thomson et al., 2025). AI-generated content is increasingly filling the pages of online reviews and articles and is potentially threatening the livelihood of online content providers located in the Global North and South. This is not to mention the impact that the digital divide has on the one-third of the world's population without access to AI in terms of economic and epistemic opportunities. Indeed, AI represents a serious problem for people whose link to the global digital economy is more precarious. There is a widespread expectation that AI will exacerbate the

concentration of income in the hands of the top 1% and widen North/South and rural/urban income gaps (Oxfam Australia, 2025, June 26; Zhang et al., 2025). This concentration of wealth bound up with the proliferation of digitalisation and AI signals the emergence of a new population of what Bauman termed superfluous, useless, and thus unwanted consumers that are no longer able to afford goods and services produced by the global economy (Bauman, 2004).

Epistemic Justice

While Kurzweil is concerned about how AI might transform societies, he is a firm believer in the power of the 'invisible hand'. That is, he thinks that corporations will make sure that their AI products are fit for purpose. Because if they are not, he argues, they are likely to be punished by potential customers and – ultimately – by shareholders (Kurzweil & Rose, 2025). Academic observers and AI models themselves are less certain about whether markets will send the right signals and thus make for an effective auto-regulation of tech corporations in the absence of what some have started to call corporate epistemic responsibility (Lamy, 2022) and hard government regulation. Corporate epistemic responsibility, they argue, is central to an era in which corporations increasingly wield public epistemic authority (Neame, 2025).

The digital economy and AI have undermined democratic processes, circumvented legislation, and pulverised human rights, are generating biased misrepresentations of minorities, are privatising and shifting our epistemological architecture, encroaching on human labour markets, posing an existential threat to millions, and are submitting many of us to an increasingly intrusive surveillance regime. Moreover, AI is exacerbating income inequality. The combined outcome of these developments on the well-being of those made superfluous will be immense and will impact at a cellular, individual, communal, and socio-economic level that will reverberate across societies for generations. AI has in store for many of us instances of serious "dignity violations" (Kar et al., 2025) and stressors that are likely to impact hundreds of millions of people around the globe, and particularly members of minority groups facing a number of intersectional stressors (e.g., black neuro-diverse men living with substance use issues). If anything, the above highlights that corporate self-regulation is myopic at best and that although AI products may be useful to some segments of society (i.e., the winners of the AI revolution), this does not mean they align with human goals more generally. Indeed, we have to find ways to mitigate the social, psychological, and ecological fallout of this 'rulers' revolt' (Edgerton, 2008), that transports us to a more unequal society.

Considering the emerging evidence canvassed in this chapter, one could arrive – not unreasonably – at the view that the development and deployment of LLM-based AI has taken place with minimal consideration for the dignity and well-being of ordinary people or the environment. This needs to change. A first step might be to redefine alignment, centring it on a framework focused on human rights and dignity. In bringing the social determinants of health to a point, Kar and

colleagues affirm that "dignity is not an idealistic concept but a biological, social, psychological, and legal imperative" (Kar et al., 2025, p. 5). Indeed, this chapter suggests that dignity also has a socio-economic and epistemological dimension. Tech corporations must take seriously their ethical responsibilities and must make human rights and dignity key alignment principles that stands for more than a rhetorical device. Dignity violations committed by AI should be systematically audited by independent bodies. Corporations should have a responsibility to repair and offer reparations and support to those affected by them. In practice, this might mean that AI revenue needs to be shared globally in order to address dignity violations. The tech corporations that collectively own most of our digital assets have been permitted to operate as if they were ordinary businesses producing goods and services that are not central to the economic, social, or political existence of most of us. As opting out of the digital economy is becoming increasingly difficult, the ownership of digital assets, including data, needs to be re-democratised and held in common.

To avert a dystopian scenario and to moderate AI's negative impact on humanity, we need to regulate, guide, and redirect the development and deployment of AI products so that they align not only with the goals of the 'winners' of the current industrial transformation. Counteracting the concentration of wealth, power, and control by elites is going to be the major challenge of this decade. In order to achieve this, we must support politicians and governments willing to develop and pass legislation to that end (e.g., strengthening intellectual property, anti-trust, privacy, and data sovereignty laws) and that create equity audit frameworks that keep track of and moderate the impact of AI products. Social safety nets rendered threadbare by decades of neglect and austerity need to be strengthened to support those dislodged by the AI revolution. We need to set up new training opportunities and courses to cater to a new generation of jobs – including new social work jobs. Moreover, we must generate political support for a general basic income, a living wage, available to those structurally unemployed that have been dislodged by the AI revolution. To generate the political will to achieve this in a time of political polarisation and fragmentation is no small feat.

Ironically, the creation of a new population of superfluous, useless consumers (Bauman, 2004) offers new opportunities for social workers. Those humans turned, in Zygmunt Bauman's terms, into the "waste product" of the most recent technological revolution are likely to feed a steady stream of new clients reliant on social work support. Whereas conventional forms of social work will face competition from AI and will be re-classified at a lower pay level, new fields of practice will open up for social workers that are experts in a profession that will increasingly incorporate digital know-how.

Concluding Thoughts

LLM-based AI has the potential to shift our epistemological architecture and, with it, the power and wealth associated with knowledge. Given this, AI harbours a tremendous potential to disrupt the social. As AI-based facial recognition, meeting

transcription tools with case note and referral functions, automated case management, and counselling chat bots are quickly becoming a mainstream component of social work practice, it is more important than ever to become aware of the magnitude of the change that we are witnessing and to gauge its political implications for a social work profession that is collaborative, anchored in human rights, and striving for social justice. Social workers need to become aware of the impact centralised control over their professional body of knowledge is having on their practice and, most importantly, on their clients and take action. Critical thinking, AI literacy, and policy advocacy skills are crucial building blocks to incorporate into social work practice alongside joining movements that seek to democratise the information produced by the digital economy and that strive for greater epistemic justice. Most importantly, however, social workers must walk alongside and advocate for the people they support to ensure that dignity violations are noticed, taken seriously, and repaired. They must become part of a renewed critical social work that identifies, reports, and monitors digital injustices and digital rights abuses and that advocates for human epistemic agency and digital justice.

References

Amladi, D. (2024, June 28). Calling for an end to unjust workplace surveillance at Amazon and Walmart. https://www.oxfamamerica.org/explore/stories/calling-for-an-end-to-unjust-workplace-surveillance-at-amazon-and-walmart/

Appel, R., McCrory, P., Tamkin, A., McCain, M., Neylon, T., & Stern, M. (2025). *The Anthropic Economic Index report: Uneven geographic and enterprise AI adoption*. https://assets.anthropic.com/m/218c82b858610fac/original/Economic-Index.pdf

Bai, H. (2022). The epistemology of machine learning. *Filosija. Sociologija*, *33*(1), 40–48.

Bauman, Z. (2004). *Wasted lifes: Modernity and its outcasts*. Polity Press.

Bitter, A. (2025, 19 Sept). Uber's CEO says robotaxis could displace drivers in 10 to 15 years and create 'a big, big social question'. *Business Insider*. https://www.businessinsider.com/uber-ceo-khosrowshahi-self-driving-car-job-loss-real-issue-2025-9

Charara, S. (2025, Feb 26). Your boss wants you back in the office. This surveillance tech could be waiting for you. *WIRED*. https://www.wired.com/story/your-boss-wants-you-back-in-the-office-this-surveillance-tech-could-be-waiting-for-you/

Coeckelbergh, M. (2023). Democracy, epistemic agency, and AI: political epistemology in tmes of artificial intelligence. *AI and Ethics*, *3*, 1341–1350. https://doi.org/10.1007/s43681-022-00239-4

Edgerton, D. (2008). *The decline and fall of the British Nation. A twentieth-century history*. Allen Lane.

Esposito, E. (2023). Does explainability require transparency? *Sociologica*, *16*(3), 17–27. https://doi.org/10.6092/issn.1971-8853/15804

EY. (2023, Dec 6). New EY research reveals the majority of US employees feel AI anxiety amid explosive adoption. *EV Americas*. https://www.ey.com/en_us/newsroom/2023/12/ey-research-shows-most-us-employees-feel-ai-anxiety

Gmyrek, P., Berg, J., Kaminski, K., Kaonopczynski, F., Ladna, A., Nafradi, B., Roslaniec, K., & Troszynski, M. (2025). *Generative AI and jobs: A refined global index of occupational exposure* (ILO Working Paper 140, Issue. https://www.ilo.org/sites/default/files/2025-05/WP140_web.pdf

Goldman Sachs. (2025, Aug 13). *How will AI affect the global workforce?* https://www.goldmansachs.com/insights/articles/how-will-ai-affect-the-global-workforce

Hadfield, L. (2022). The hidden cost of the gig economy: Rights for injured gig workers. *Precedent (Australian Lawyers Alliance)*, *November/December* (173), 31–35. https://www.austlii.edu.au/au/journals/PrecedentAULA/2022/65.pdf

HAI. (2025). *Artificial intelligence index report 2025*. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

Hodgson, D., Gair, S., & Watts, L. (2026). Resisting the enchantment of LLMs: Ethical implications for social work practice, research and education – a case study. In G. Ottmann & C. Noble (Eds.), *AI and the disruption of welfare: Challenges for social work education and practice* (pp. 139–152). Routledge.

Hutson, M. (2025). AI models that lie cheat and plot murder: Just how risky are LLMs? *Nature*, *646*(9 October), 272–275.

IASSW. (2014). *Global definition of the social work profession*. International Association of Schools of Social Work. Retrieved 16/06 from https://www.iassw-aiets.org/global-definition-of-social-work-review-of-the-global-definition/

IFSW and IASSW. (2014). *Global definition of social work*. https://ifsw.org/get-involved/global-definition-of-social-work/

ITU. (2021). *Measuring digital development: Facts and Figures 2021*. https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf

Kafka, P. (2025, 29 Aug). Donal Trump once threatened to jail Mark Zuckerberg. Now he's his ally. *Business Insider*. https://www.businessinsider.com/donald-trump-mark-zuckerberg-meeting-allies-meta-2025-8

Kar, A., Ivbijaro, G., & Rabheru, K. (2025). Dignity is the methodP: ethnic minority mental health, strucural harm, and the constellation model. *Academia Mental Health & Well-Being*, *2*(4), 1–6. https://doi.org//10.20935/MHealthWellB7932

Kemp, L. (2025). *Goliath's course: The history and future of societal collapse*. Viking Penguin Random House.

Kurzweil, R. (2024). *The singularity is nearer: When we merge with AI*. The Bodly Head, Vintage, Penguine Random House.

Kurzweil, R., & Rose, D. S. (2025, May 18). *The sigularity is nearer*. The 92nd Street Y. https://www.youtube.com/watch?v=7oAlD3lMNXo

Lamy, E. (2022). Epistemic responsibility in business: An integrative framework for an epistemic ethics. *Journal of Business Ethics*, *183*, 1–14. https://doi.org/10.1007/s10551-022-05078-1

Lee, T. L., Tapia, M., Lima Aranzaes, C., Sapre, S. R., & Bustamante, A. R. (2024). The militarization of employment relations: Racialized surveillance and worker control in Amazon fulfillment centres. *Work and Occupations*, 1–38. https://doi.org/10.1177/07308884241292733

Leiter, B. (2024). Free speech on the internet: The crisis of epistemic authority. *Daedalus*, *153*(3), 91–104. https://www.jstor.org/stable/48784943

Liu, L., Wang, Y., Zhao, H., Kong, S., Teng, Y., Li, C., & Wang, Y. (2025). Reflection-bench: evaluating epistemic agency in large language models. *42nd International Conference on Machine Learning*, Vancouver, Canada.

McQuillan, D. (2019). Non-fascist AI. In M. Hlavajova & W. Maas (Eds.), *Propositions for non-fascist living: Tentative and urgent*. BAK basis voor actuelle kunst and MIT Press. https://osf.io/preprints/socarxiv/b64sw_v1

McQuillan, D. (2022). *Resisting AI: An anti-Fascist approach to artificial intelligence*. Bristol University Press.

Messeri, L., & Crockett, M. J. (2023). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*. https://doi.org/10.1038/s41586-024-07146-0

Mollnick, E. (2024). *Co-intelligence: Living and working with AI*. Random House UK.

Mueller, V. C. (2021). Ethics of artificial intelligence. In A. Elliott (Ed.), *The Routledge socila sicnece handbook of AI* (pp. 122–137). Routledge.

Muldoon, J., & Wu, B. A. (2023). Artificial intelligence in the colonial matrix of power. *Philosophy & Technology*, *36*(4), 80. https://doi.org/10.1007/s13347-023-00687-8

Neame, L. (2025). Facts, arte-facts, and fabrications: The crisis of public epistemic authority. *Political Theory*, 1–38. https://doi.org/10.1177/00905917251321274

Oxfam Australia. (2025, June 26). New wealth of top 1% surges by over %52.4 trillion since 2015 – enough to end poverty 22 times over, as Oxfam warns of global development 'abysmally off track' ahead of crunch talks. https://media.oxfam.org.au/2025/06/new-wealth-of-top-1-surges-by-over-52-4-trillion-since-2015-enough-to-end-poverty-22-times-over-as-oxfam-warns-global-development-abysmally-off-track-ahead-of-crunch-talk/

Petrosyan, A. (2025, April 1). Worldwide digital population 2025. *Statistica*. https://www.statista.com/statistics/617136/digital-population-worldwide/

Rani, U., & Pons-Vignon, N. (2025). Introduction: Power relations in the digital economy. *New Political Economy*, *30*(3), 313–324.

Suzuki, S. (2025). Delegating tasks to AI encourages cheating. *Nature*, *646*(2 October), 42–43.

Thomson, T., Nagappa, A., & Weinbrand, S. (2025). *AI overviews have transformed Google search. Here's how they work – And how to opt out.* https://theconversation.com/ai-overviews-have-transformed-google-search-heres-how-they-work-and-how-to-opt-out-258282

Venkit, P. N., Li, J., Zhou, Y., Rajtmajer, S., & Wilson, S. (2025). A tale of two identities: An ethical audit of human and AI-crafted personas. arXiv, 2505.07850v1. https://doi.org/10.48550/arXiv.2505.07850

Venkit, P. N., Zhou, Y., Mao, Y., Laban, P., Huang, K.-H., & Wu, C.-S. (2025). DeepTRACE: auditing deep research AI systems for tracking reliability across citations and evidence. arXIiv, 2509.04499v1. https://doi.org/10.48550/arXiv.2509.04499

Ylä-Anttila, T. (2018). Populist knowledge: 'Post-truth' repertoires of contesting epistemic authorities. *European Journal of Cultural and Political Sociology*, *5*(4), 356–388. https://doi.org/10.1080/23254823.2017.1414620

Zhang, W., Han, J., Işık, C., Song, J., Liu, X., Dembińska, I., & Radulesu, M. (2025). Inequality in the digital economy: The impact of artificial intelligence on income gap—An empirical analysis based on county-level data from China. *Cities*, *166*, 106280. https://doi.org/10.1016/j.cities.2025.106280

Zuboff, S. (2019). *The age of surveillance capitalism: Fight for a human future at the new frontier of power*. Profile Books.

3 Social Work in the Hypermodern

Advancing Humanism in the Age of Smart Machines

David P. Moxley

Smart Machines as an Emergent Social Problem

Although the disappearance of employment in low-income communities was documented by Wilson (1997) over 25 years ago, the talk is now about the erosion of employment among the educated and professional classes in the United States because of the rapid implementation of Artificial Intelligence (AI). Governments, such as Britain and states like California, are already setting limits on AI, anticipating that the technology itself is a disruptor requiring careful oversight, even though it is difficult to craft policy given the uncertainty of the technology and its now rapid dissemination (Harding, 2024).

The real test of this technology is whether it can solve a broad array of problems in varying environments, imbuing the technology with capacities for general intelligence (Togelius, 2024). For humans, intelligence is likely environmentally specific, affirming the importance of specialization in modern societies. But it is exactly in these specialized domains that AI challenges human beings because of its expansive memories and capabilities for rapid retrieval of information, with implications for replacing humans in professional fields. Still, although we may arrive at a place in which human beings come to see AI as augmenting the skills of highly educated and trained professionals, such as specialized physicians, the technology will likely replace workers whose expertise involves the routine manipulation of symbols and analysis, such as accountants or even attorneys. A paradox is forming in which society in most domains, including the humanities and the arts, is questioning expertise and knowledge, and not a few politicians are dismissing the legitimacy of established knowledge – the kind one gets from years of specialized education and training (Nichols, 2024). Yet, AI as a potential replacement for human analysis, interpretation, and decision-making is now taking root. One can argue that AI cannot achieve the subtlety of human intelligence, given how human beings make meaning, particularly within cultures, but my observations are that the technology can very much replace the kind of routine work professionals undertake across a diversity of fields.

Advanced Technologies as Inexorable Forces in Social Disruption

AI in its contemporary context has been long in coming, building on some 200 years of science devoted to calculation (Pasquinelli, 2023). Thanks to the expansion of

DOI: 10.4324/9781003627920-3
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

relatively inexpensive memory, storage, and retrieval, use of AI is already changing profoundly the jobs of journalists, commercial airline pilots, lawyers, accountants, performance artists, like musicians and actors, and healthcare providers. American society is moving toward the widespread use of automation, which made its advent in science fiction over 150 years ago and gained ground as manufacturing in the United States and other countries automated agriculture and manufacturing first through mechanized processes and then through digital ones.

In a seminal study of changes to manufacturing using smart machines, Zuboff (1988) observes that technology would shape human organization, which now in 2025 holds significance as human beings interact with digital machines with considerable memory, that rival if not exceed the capabilities of workers in almost all domains of society. The advent of big data analytics and predictive software is further altering the fabric of professional work. What was once seen as the use of technologies in the augmentation of work undertaken by humans, such as in manufacturing, is now a technology that can replace humans that perform abstract functions.

The substitution effects of AI for abstract labour, the kind professionally trained and highly educated human beings have undertaken, hold considerable implications, some would say negative implications, for restructuring the workforce in the professional domain (Susskind & Susskind, 2015). In the age of smart machines, what does it mean to be a professional? Related here are other critical questions: What does it mean to be an expert (Susskind & Susskind, 2015)? And will society take expertise seriously when smart machines are dominant and access to information in automated forms is so abundant (Nichols, 2024)? After all, the rapid development and diffusion of AI have outstripped historically that of other technologies (Floridi et al., 2018). AI is now evolving so rapidly that it is reverberating broadly through society.

Digitalized machines, some with robotic capabilities, became even more prevalent in industry during the 1980s when the American automobile industry shifted the paradigm of manufacturing following the lead of Japan in the use of automation. Along with the use of robots in manufacturing, high-speed data-analytic methods and algorithmic analysis began supplementing human decision-making, especially in manufacturing and business environments. In the 1990s, robots became pervasive in manufacturing. Socially assistive robots are coming soon, in which machines can assume the responsibilities of direct care workers whose patients require physical support and assistance with mobility and movement. Socially assistive technologies will likely become pervasive in the education of children and adults (Yale University, 2025). Personalized smart technologies can provide social support for individuals who need emotional reassurance, and those sources of support can help people remember when to take medication, engage in health promotion practices, adhere to diets, and control appliances in the home to ensure the safety of inhabitants.

Jobs immediately at risk are those that involve routine cognitive procedures amenable to digital programming, including higher-level cognitive tasks (Winthrop & McGivney, 2016). Workers with highly skilled positions may integrate AI

into their work as they adapt such technologies as assistants. But these assistants may have their own limitations, such as the inability to create mental models of the world (Bender & Hanna, 2025). Coeckelbergh (2020) questions the capability of AI in the creation of meaning. The technologies may lack the capability of intentionality in which humans make meaning and then assign that meaning to machines (Coeckelbergh, 2020). Smart machines, as Boyle (2024) warns, may require humans and their institutions to recast the line between human functioning and smart machines, as AI raises questions about its rights, its moral standing, and its own autonomy. Yet, the interactions between human intelligence and AI hold some promise in advancing the distinctive qualities of human thinking, particularly in how culture shapes such thinking (Brockman, 2019). However, the technology may exclude those who simply cannot afford its use, raising questions of social justice in access to powerful computing capabilities (Crawford, 2021). Advanced technologies create, as Pepi (2025a) emphasizes, considerable inequality.

Writing in the *Washington Post*, columnist David Ignatius observed in 2016 the anger welling up among working-class people whose income, jobs, and opportunities have eroded over decades of corporate and governmental neglect and disinvestment. Advanced technology has already restructured the employment of those who drive trucks or taxis for a living, arrange travel, provide routine healthcare, and calculate for a living, like bookkeepers and accountants. The advent of drones, once seen principally as a military technology, is now appearing in the transportation industry as self-navigating and self-piloting technologies come to replace human operators. Technology is advancing surveillance of personal lives, on-line presence, community involvement, consumption, and speech with an ease once thought improbable introducing significant concerns about liberty and constitutional rights to privacy (Zuboff, 2019).

Smart or intelligent technologies can either augment or replace professionals engaged in abstract work or technicians who rely on their judgment to maintain or operate machines. The potential displacement of huge swaths of the American workforce is not coming, it is happening, and it is gaining ground. The movement for replacement may quicken as the American workforce changes in composition, the demographics of the workforce shifts, and baby boomers retire in mass. Ignatius argues for policies that protect economic security (as opposed to job security) like universal income, massive public employment projects, and the support of consumption. Pepi (2025b) identifies the Digital New Deal as a necessary policy promoting equity involving taxing of AI companies that create counterfeit content of human originals. In his white paper, Pepi notes that

> (in) the hands of Silicon Valley, the primary export of AI firms has come to be known as AI Slop—AI generated amalgamation of the runoff from billions of harvested data points, served up as a low-effort synthesis paste by platforms (and for platforms).

The legal vulnerability of AI is gaining ground as major AI corporations face copyright lawsuits for the pirating of copyrighted material (Bloomberg Law AI, 2025a).

Physical and cognitive technologies can displace human labour profoundly. The increasing presence of advanced technologies operating with AI could reduce wages or perpetuate stagnant wages. A displacement of semi-professional and professional workers could dwarf the displacement of agricultural workers the United States experienced in which years ago, as a result of mechanization, field hands became factory workers in the growing cities of the United States. Some economists expect that the emergence of new jobs will compensate for the displacement of workers as a result of a growing use of technology. But still, those most vulnerable may experience high rates of displacement as they have experienced during any transition or upheaval in technology. New populations may join the ranks of the vulnerable, further stretching an already overwhelmed, inadequate, and obsolete social safety net.

Populism is growing across the globe, but it is not necessarily progressive. A high technology economy, one dominated by machines, and synthetic intelligence operating within those machines, but now ones that can be described as smart, can marginalize workers in manufacturing, food preparation, and health care shifting the global workforce in uncertain ways. Ignatius writes that presently "5% of occupations can be automated, but 60 percent of occupations could soon see machines doing 30 percent or more of the work." The human worker who holds a smart manufacturing job, a relatively high paying albeit hourly wage job, may soon evaporate as machines continue to incorporate routine physical and cognitive skills that were once so very abundant in the workforce. Indicators point to a decline in workforce participation among those who are relatively young workers, and there is some evidence too for declining work hours (Meyerson, 2014). The recovery from the Great Recession of 2008 was stubborn principally because large employers replaced human workers with technology, a typical outcome of many economic recoveries in advanced economies. Today, in 2025, advanced technologies can handle complex tasks once within the purview of trained professionals, such as coding, more productively, and faster than humans.

In the past, workers with so-called low-level skills, ones who used brawn in their work, were displaced by machines, particularly in the agricultural sector, and then in low-level manufacturing. Now, middle-tier occupations, and the jobs those encompass, may experience substitution of smart machines with the ability to learn, with capabilities for deep autonomous learning, potentially marginalizing smart humans and creating the surplus populations of the near future. Machine learning is now well positioned to make decisions independent of human beings (Ananthaswamy, 2025).

The erosion of middle-tier jobs will likely perpetuate the political crisis in the United States, and in other western countries, as political leaders have few options for advancing mass employment of people who expect good wages. These wages are needed given the inflation of 2025 as households face inexorable price increases in the core areas of housing, energy, food, and health. Affordability has emerged in 2025 in signifying the inflationary forces consumers face although politicians may ignore the realities these consumers must address in their daily lives.

As a result of the introduction of a new generation of smart machines, populism will likely grow, involving mass movements to protect jobs, or to retaliate against

corporations substituting advanced technologies for human labour, and its negative consequences will echo throughout societies. Desperation among people may grow resulting in populist authoritarian leaders offering suspect solutions while they use the very technologies they rail against for political advantage, particularly in manipulating information.

The continued displacement of labour may create interrelated social issues involving crime, homelessness, insurance, and the creation of so-called portfolio jobs. Tax structures may shift in which sporadic employment or unemployment falls short of supporting local communities that depend on property taxes for the support of local services, such as firefighting, policing, and education. Those working will likely need to shoulder a higher proportion of local taxes requiring them to willingly accept these responsibilities, albeit grudgingly.

Given how the United States structures its tax system, more responsibility will fall on homeowners who shoulder considerable burden through property taxes to finance local government expenditures. With retirement accelerating among members of the post–World War II Baby Boomers, entitlements will continue to become a politicized issue at national levels. The Trump administration's restructuring of the federal budget will likely mean fewer resources available for local communities. Displacement of well-paid federal employees and their substitution through AI is the wave of the near future.

The Hypermodern

The age of the smart machine and AI is upon us. And this age is hypermodern. The modern brought rationality, systematic approaches to work, the importance of professional work, the expansion of higher education, and the sanctity of professional judgment and discretion. In the modern era, public higher education was a touchstone offering new opportunities for those who completed professional or even liberal arts courses of study. The Servicemen's Readjustment Act of 1944 in the United States (colloquially known as the GI Bill) ushered in the advancement of higher education. The coming of age of the Baby Boomers in the 1960s further expanded higher education considerably. Given demographic shifts and the political attack on higher education, public universities are under assault. State budgets cannot sustain public higher education at previous levels and inflation outstrips other sectors as colleges and universities build boutique campuses to appeal to students who value a broader base experience than what academic preparation can produce alone.

Higher education is losing its lustre as the principal opportunity structure of American society. Demographic shifts do not favour universities, particularly ones founded in the 19th century on a model of liberal arts education. Automation and digitalization anticipate the need for fewer educators. Decreases in federal support of science, the social sciences, health, and the arts and humanities may sideline existing faculty as demand decreases for specially trained faculty members whose knowledge base may become increasingly redundant or simply unwanted. The Trump administration's defunding of higher education and research sidelined many faculty structurally changing the careers of scientists, humanists, and health

care faculty who have depended on the availability of federally funding of research prevalent since the inception of the Cold War.

The hypermodern is about rapidly changing information or digital technology creating at least two kinds of social change in relationship to work and the workforce: it alters current membership characteristics of the workforce in the sense of who can participate in work and what those workers do, and second, it anticipates a future in which human participation in employment falls, perhaps precipitously. In the age of the hypermodern, smart machines can replace smart people. As a result, professions and the character of professionalism are undergoing considerable modification or transformation (Susskind & Susskind, 2015).

Technology, however, is never a panacea for social ills. They can create new ones and exacerbate old ones. Writing in the *Guardian*, journalist Adrienne Matei (2025) underscores the idea of hypernormalisation who, taking this concept from a deteriorating Soviet Union, characterizes contemporary United States as emerging from the country's deterioration of its democratic norms, dysfunction of governing systems, in which major social institutions are breaking, incompetence of political leadership and the resulting disaffection among common citizens signal collectively that a new normal is forming. Considerable fear, anxiety, denial, and withdrawal characterize this new normal, according to Matei, as people witness almost daily the decay of systems they need and on which they have come to rely if not expect.

The sweeping changes brought on by the new smart technologies contribute to this sense of dysfunction in hypernormalisation and what I call the hypermodern. The worry that smart machines will displace workers is unfounded according to critics (Rampell, 2015). But this argument is based on the idea that those who are most qualified will find employment in other sectors or jobs in newly emerging industries that AI could possibly open up. Unfortunately, those most vulnerable to technological shifts will likely suffer the most. The United States has witnessed this in the marginalization of rural communities, in reduced mobility for working-class and middle-class workers, wage stagnation, and the shifting of manufacturing to offshore locations. The promise of globalization, a key assertion of the Clinton Administration, did not create jobs for those who are vulnerable to even marginal shifts in the economy. The North American Free Trade Agreement (NAFTA), and its open markets, particularly in the United States' trade with Canada and Mexico have failed the working class, while other classes have prospered.

Professionals themselves will find their skills increasingly accessible through AI or machine intelligence. Susskind and Susskind (2015) offer a dystopian scenario of the transformation of professionalism. They offer a classification for how workers will accommodate the displacement of professionalism. Within this scenario nonprofessional or quasi-professional roles will do the work and jobs that smart technology drives. These are the assistants to smart machines. Paraprofessional roles come about through the necessity of such machine service activities, and those roles demand less investment than what traditional professional roles would otherwise demand. Bean (2024) suggests that new kinds of workforce innovations, such as digital apprenticeships, can equip workers with the skills (and insight)

to collaborate with smart machines in effective and meaningful ways. Lifelong education may emerge as people move in and out of the workforce to augment or refine their skills and corresponding knowledge as local colleges and even universities take on more responsibilities for workforce education, development, and transitions.

Widespread access not only to information systems, but to knowledge forming analytic smart technologies can increase the scope of the availability of what was once professional knowledge but is now part of the information cultures of specific groups, such as legal practices, and educational systems. Lawyers may make use of cases that renegade AI systems compose, raising questions about the trustworthiness of machine-generated algorithms. There are calls for the establishment of an AI incident database to document the malfunctioning of AI illustrating that the technology is fallible (Bloomberg Law AI, 2025b). Professional roles will likely change in the near future as technology replaces or substitutes what was once sacred knowledge kept and used by professionals. Machine-generated knowledge, however, may be as imperfect as the knowledge human professionals create and use.

Access to specialized knowledge bases through AI can alter the dialogue and interactions between professionals and clients, such as when physicians interact with their patients who understand the state of practice in certain treatment areas because they prepared themselves for their visit with their physician using AI. AI may shift health care from specialized physicians to less trained professionals, like nurses, who can access specialized relevant practice knowledge through digital systems. Many areas such as rural and remote places without the availability of physicians may replace those with other practitioners, like Advanced Nurse Practitioners further accelerating trends that are already occurring. Still, the propagation of telehealth interacting with AI can augment health care resources fostering productive interactions between patients in their own homes and remote health care providers as evidence-based practices standardize medical care. Yet, such interactions require adequate digital technologies, absent in many rural or remote areas of the globe.

Two prominent strategies for the adaptation of professionals to AI involve human-machine collaboration in the production of algorithms that machines need to sustain their intelligence and improve performance, or when highly trained professionals either individually or in groups outperform the intelligence of smart technologies. Either strategy narrows the occupational niches that have been available in the past, those demanding some amount or high levels of higher education as an entry credential for careers that society has needed for its own development. These occupational niches required the operationalisation of highly abstract modes of cognition, something the advocates of AI say this technology can mimic. What was once the purview of professionals can be automated, or at least this is the argument of AI advocates.

Economists may equate progress with substitution effects. Augmentation of workers can increase productivity, but substitution can eliminate human factors altogether. The general effect is a more efficient production and automated use of

knowledge within society, but workforce participation among humans can potentially plummet with deflation akin to those years anticipating the Great Depression of 1929 or the Great Recession of 2008. The hypermodern may set in motion serious negative economic consequences, one's that policy makers may readily accept as the cost of societal transition to a new economy.

Truncation of Work in Hypermodernity

The hypermodern age is now here, and it holds important if not critical implications for social work and all of its disciplines, clinical practice, group work, community practice, and social administration. The unavailability of work is central to almost all of the social problems social work addresses but so is the unavailability of good work that carries with it liveable wages and good benefits. Most of the social problems social work addresses involves the challenge of work—too much work, not enough work, the unavailability of work, and poor work conditions. The deprivation of work interacts with the availability of health care insurance, adequacy of retirement income, poverty, substance use, and family disruption.

Displacement of workers creates major social issues within communities. Social workers are likely sensitive to the central role employment, or its absence, plays in the lives of the people they assist compromising the safety and quality of life of whole neighbourhoods. Despite the cultural centrality of work in the United States, there is no guarantee of employment, and employment itself is not a social right. Indeed, that work is not really a social right in the United States reveals the limitations of the United States Constitution in its glaring omission of a bill of social rights (Sunstein, 2004).

Smart technology interacts with the age-enduring problem of ensuring the availability of good work for all people, work that pays adequate wages but, more importantly, liveable wages. The linkage of work with benefits, like health care, and retirement income, makes the American system of employment problematic, especially for those with limited workforce participation. Substantial cuts in Medicaid as an aim of the Trump administration's "Big Beautiful Bill" exacerbates already serious limitations in the American social safety net. The availability of jobs for many people is central to their social and personal development but the recognition of the human value of work does not factor into policy deliberation the importance and purpose of work in American society. Given the cultural, social, psychological, health, economic, and physical importance of work to human beings, can a constitutional amendment guaranteeing work as a right serve as an outcome of the current populism? In the hypermodern, can work as culturally central but not guaranteed become work as a right?

With work uncertain or even in jeopardy for people in the age of the hypermodern, can society address the potential disappearance of gainful employment by combining new forms of education and training (that can postpone entry into the workforce or assist people who are in transition while equipping them with tangible and market-relevant skills) with work as a right, universal basic income, and universal health care? A coalition of working class and professional groups, those

whose work lives the hypermodern may threaten the most, may seek such policy innovations, especially when they demand an adequate basic income structure that reinvents the redistribution policies of post–World War II Europe and the decades of the 1960s and 1970s in the United States. The hypermodern with its focus on the use of advanced technologies can spur coalition social movements the membership of which cuts across economic classes of society. Perhaps what emerges is a unified set of policies focusing on work, income, health, food, housing, and energy, ones that frame rights to those resources well-being requires.

Regulation of entry into the workforce is not foreign in the history of the United States. Anxious about the effects of demobilized soldiers returning en masse to the United States after World War II, and the sheer volume of those veterans who would be looking for employment, the Truman Administration postponed troop movements back to the country. However, when they did arrive, male veterans displaced women who worked at highly skilled jobs during the war and who came to enjoy the positive benefits of working. Even relatively progressive legislation, like the GI Bill, sought to reduce demands on employment as male veterans entered academic programs in lieu of immediate employment. Yet, in this situation higher education held the promise of the entry of veterans into the professions, which it may be unable to offer to a new generation of students. Indeed, higher education may move away from the education of students for the professions or the preparation of students for a diverse workforce through liberal arts education in favour of offering highly technical education that supports automation and AI. The shelf life of this education may be short lived as technological innovations and their substitution effects require lifelong learning (Dede & Roberts, 2020). Advanced technologies may dampen the lustre of education and reduce its promise in offering labour a competitive advantage over what advanced technologies can achieve. The wholesale restructuring of higher education is likely to come.

But this policy configuration may be insufficient to address the broader cultural and institutional consequences of the hypermodern. A cultural shift may be necessary in which work is less central to the identities of individuals. In the age of the hypermodern, people may work less, completing a work week in three days, share jobs, and be deployed by their employers to work on community projects, like supporting elders, engaging in child and youth development, or undertaking ecological restoration. The advent of compulsory national service for multiple age groups may occur as countries seek to offset the economic and social displacement of various age groups. This could be yet another tactic for forestalling the movement of groups into the workforce or into the ranks of the under- or un-employed.

If the hypermodern threatens the availability of work yet another issue is under-employment. This possibility indicates that the substitution of smart technology for human physical and cognitive labour may create an abundance of talented human capital to address social needs. Cognitive surplus may accumulate in communities and make available creative solutions to what is now seemingly intractable problems (Shirky, 2010). It is here that social work in the hypermodern becomes relevant.

Ethical Considerations

Most social workers will likely endorse the idea that work is essential to the human experience, and something would wither without the kind of productive activity we call work. Human beings are self-reflective and possess considerable potential for framing and creating personal as well as group identities. The value of work may be one that multiple cultures share and comes to serve as a critical referent point in the manner in which people relate to one another. Imagine meeting someone for the first time, and you hear yourself asking: "What do you do for a living?" The person comes to use a particular label to identify oneself. Freud came to see happiness as involving love and the self-fulfilling features of work. Marx spoke of the "essence of the species" to underscore the importance if not centrality of work and its production of value in the human experience.

The Luddite movement in early 19th-century England came to define people who were against technology and its displacement effects as machines took over in the British textile industry and as textile workers fought the use of power looms. Yet, the Luddite movement tells us something about the centrality of work to human development and in affirming the importance of craft. The Dutch workers who jammed their sabots (wooden shoes) into machines to stop production as a form of protest produced the term sabotage, a word taking on nefarious connotations almost equivalent with the contemporary use of the term terrorist.

If work is central to the human experience, and it offers meaning to people individually and within group or community life, what is its ethical importance within a given society? Does a society forego the economic benefits of substitution? Does it speed up its use so it can deploy workers to other productive roles? Does technology become a more important end than humanism, the preservation of human merit and worth, and the protection of dignity?

Societies have faced these policy questions in the past. The mechanization of agriculture is a good example of technology substituting for human labour. Such substitution pushed whole groups of labourers off farms, eliminated or obviated tenant labour, and urbanized the proletariat who were once agricultural workers and then who became factory labourers. Still, the automation movement in the automobile industry reshaped entire manufacturing paradigms, and through precision tooling, loading machines, and smart technologies further reshaped work and the human worker, who became an attendant and not necessarily a creator.

This history leads to an ethical question about the merits of human beings. What is the end of work, to use a phrase with double meaning? Is the end the elimination of human work as we know it? Is the end of work the elaboration of human identity and the pursuit of meaning? The ethical consequentialist may urge us to focus on the results. Positive results will be visible in happier people who have better things to do than work. Negative results will be visible in wholesale displacement resulting in human despair and a future of economic deflation. However, humanism may offer the most trenchant response: human beings find fulfillment through lives of dignity (Flynn, 2000). Good work and good paying work reinforce human dignity found in the identity of individuals, group life, and culture (Kateb, 2011).

The availability of work and liveable wages is a cornerstone of such dignity (Moxley, 2025). But technological advancement, as Schon (1967) observed decades ago, serves as an ethic of its own within society. What is the rationale for restraining technological advance in the face of negative consequences particular groups may experience? Governments and politics need to address Schon's observation since AI has its own path of transformation, perhaps one isolated from the greater good a society may seek, at least if it values all of its members.

A cautious policy inherent in "wait and see" will slow down the process of substitution. A policy of unbridled entrepreneurialism may favour the wholesale liberation of new technologies that can further corporatism without responsibility in a society. Those who subscribe to an "anything goes" stance may simply argue for the liberation of smart technologies as a social experiment of sorts. What would we do as smart technologies have the knowledge base if not the probabilistic judgment of human beings and, as a result, the society deploys them in mass within business, military, education, health care, communications, and other pivotal sectors of society? Can cultures, including legal and regulatory ones, as well as institutional ones, evolve quickly and with agility to keep pace with this unprecedented level of disruption?

Although ethics may have much to say about these policy stances, market forces may resolve the questions about the use of smart technologies more quickly than the policy process can handle. Still contemporary neoliberal ideology argues that most solutions are best left to markets, which prizes what Gabriel (2022) refers to as economism that assigns considerable power to economic decision-making. Economic nationalism may offer a simple and unambiguous solution when politicians assert that a country will pursue AI or any transformational technology in an unbridled way prioritizing national competitiveness and security interests (and perhaps their own) over any other end. Narrow group interests will likely prevail over a broader set of interests (Fukuyama, 2018).

Some governments, however, have enforced limitations in biotechnologies with decision makers assigning considerable risk to the unbridled use of such technologies in disrupting natural and social ecologies. Regulation of genomic technologies has tempered the appetites of its advocates, at least in the United States, although their use grows in countries that do not regulate biotechnologies. Biotechnologies create considerable existential threat, but so does AI.

Is such regulation relevant to smart technology and its potential effects on the availability of good work for human beings? There could be very good ethical and policy considerations for limiting smart technologies to reducing or even eliminating hazardous industries and jobs, like mining, meat packing, roofing, and some forms of clerical work in which repetitive motion can damage the limbs, and compromise body geometry of human workers. Still, many take such jobs to make a living, and so they accept risky employment.

Social workers could evoke an ethic of consequentialism in assessing the upside and downside of AI. If the technology can displace whole communities, and whole groups of workers, and the work resulting in displacement degrades the well-being of humans, regulation can slow the process of substitution, but so can

social innovation designed to co-opt advanced technologies. Slow moving controls can serve as a means of slowing hot moving technologies. The political appetite for such regulation may be absent opening the necessity of social action in addressing the problems advanced technologies can create.

Four Scenarios Framing Social Work in Hypermodernity

There are at least four scenarios social work can consider in framing its response to the hypermodern. These scenarios reflect pure types although their blending is more likely in the near future. Each scenario suggests that social work will engage technology as a social issue within society, something the profession may be reluctant to address.

Scenario 1: The first one is the acceptance of substitution. Here, as with any social problem social work addresses, the profession engages the negative consequences of societal change. In many ways, social work is a profession that embraces the widespread social change new technologies, demographic changes, and world events can induce as it comes to address negative consequences of social change at multiple levels, such as the personal, group, community, and societal levels. Within the substitution scenario, social work engages negative consequences as people who are displaced by technology experience disruption in their personal, family, and work lives. The profession should expect that substitution will come to form a significant social problem disrupting the lives of a mass of people because of institutional malfunctioning, a classic property of any social issue. Social work may bring this problem into a clinical framework in which the profession and its clinical practitioners address the cognitive, emotional, and social disruption people experience. Rapid displacement of people may likely fill the intakes of many clinical social workers as the struggle for dignity brings depression, anxiety, and existential worry.

Scenario 2: There is a tremendously dark scenario in which AI and those who command it undermine if not intentionally violate democratic processes, collective structures, institutions, information, and privacy. Reports of this scenario dominate news portals documenting the use of bots to distort voting, the employment of deep fakes, and the intentional distribution of misinformation and disinformation. Like any technology with a track record of undermining public integrity, there will be backlashes to this manipulation, and already, such as in Europe, regulatory bodies are responding (European Commission, 2021). Can they respond fast enough? Can they address the biases that AI already incorporates? Can society address the vulnerability of those who are susceptible to such manipulation? These questions are more political than they are technological.

AI may elevate the capabilities of police in profiling people of minority status further reinforcing inequality (Eubanks, 2018). And algorithms can support an already virulent racism (Noble, 2018). Surveillance of people may gain ground through the use of technology that can detect deviance, violation of social norms, and engagement in dissent (Zuboff, 2019) with considerable investment in advanced technologies for the surveillance of immigrants and undocumented

people in the United States (Politico, 2025). The Trump administration's loosening regulations guiding surveillance anticipates broader use for social control purposes (Politico, 2025). There is a convergence here between state power and technology power (Kingsnorth, 2025) amplifying traditional power asymmetries within society.

Regulating technology has always been a challenge for governments and industries (Crawford, 2021). The regulatory possibilities raise not just economic issues, but also social development ones as society seeks to temper the humanitarian fallout of AI or any technology emerging from it (Walther, 2020). It will likely call for what Walther (2021) identifies as adaptive leadership for the purposes of social development. Social development focuses on the capability (and willingness) of society to raise people out of poverty above thresholds where well-being results from policies that focus on the development of whole communities in which all people can thrive (Walther, 2021).

Scenario 3: Standing alongside the substitution scenario is enhancement. Here people, governments, and employers may accelerate their use of AI and its smart machines creating new knowledge, skill, and performance requirements for job holders. The two scenarios of substitution and enhancement may together create somewhat of a paradox as people are displaced from existing jobs and as people modify their jobs to use smart technologies.

The enhancement scenario will likely create challenges for social work. Here social workers may increase their involvement in workforce development, technical education, and support for personal development as some workers may enjoy more leisure time. Social workers may find themselves working closely with higher education as it makes the turn into new institutional forms of lifelong, workforce, workplace, and technical education (Dede & Richards, 2020).

Enhancement may also work in unexpected ways. AI holds possibilities for improving people's quality of life through the formation of co-intelligence, the collaboration between human beings and technology (Mollick, 2024). Enhancement as policy can align with social work values in helping people reduce the impact of disability through technologies that improve human freedom and emancipation. AI can offer navigation systems for people with low vision or other sensory differences helping people to move about their communities more safely and effectively. It can rapidly translate content into multiple languages, protect people from falls, prevent accidents, facilitate transportation, and augment learning for people who can benefit from novel instructional systems. AI holds promise for enhancing human memory, creating reminder systems, and introducing novel forms of entertainment enabling people to advance their quality of day. Technologies that improve the well-being and functioning of people who face challenges as a result of disability may come into the mainstream helping everyone function better or perhaps with less effort. AI can advance medication discovery and improve food production. It offers considerable promise for enhancing the human experience.

Scenario 4: Bringing technologies into group life and social support may expand the role and influence of social work in facilitating human development, which implicates the fourth strategy: co-optation. A co-optation strategy means that social

work will purposefully engage technology to improve the well-being of people who can experience marginalization. Co-optation of AI means that social workers would purposefully pursue a strategy of availability, access, and use for the purposes of social advancement. Even under conditions of substitution in which there is a possibility of broad displacement of workers, social workers can collaborate with those who experience displacement in the creation of new opportunities and productive activities found in creative, inventive, and innovative social enterprises, ones that incorporate AI (Smith et al., 2020).

Some of these enterprises will engage technology in premodern ways in which small groups form highly collaborative and supportive efforts in the arts, crafts, performance and in infrastructure development, including food production, housing, and child and elder care. Connections between craft and technology are gaining ground in the hypermodern, perhaps as a counter movement or a form of resistance. Social innovation can emerge within this scenario as the human use of technology becomes more salient, echoing Wiener's (1950) landmark but long-forgotten plea for the "human use of human beings" found in his advocacy for the use of technology to release people from repetitive work so they could invest their gifts in the creative aspects of their jobs. In this scenario, societies may harness AI to advance creativity within the workforce.

Within this scenario, social work advocates for inclusion through the involvement of members of marginalized communities and groups in the design and use of AI, involving yet another tactic of co-optation of a technology that is now more in the hands of centralized elites than it is in the hands of those experiencing marginalization (Mohamed et al., 2020). This tactic can amplify the importance of creating cultures of AI in which those who have experienced exclusion can participate not only in the use of the technology itself but in aligning the technology with social development aims that improve people's well-being (Vinuesa et al., 2020).

Social Work Strategy in Hypermodernity

What is social work in the hypermodern? Social workers could be a central force in redefining the cultural aspects of the hypermodern. Political social work may grow in importance as social workers use group work to foster public participation in the political life of communities, with political activism growing among people who have substantial surplus time, abundant cognitive assets, and considerable untapped physical energy. My use of surplus implicates the time, cognition, know-how, and energy that work fails to absorb.

In the premodern and modern, those aspects of human resources were readily absorbed by work. The hypermodern may liberate them. Social workers may create new avenues for education and personal development of those who are searching for alternatives to work as it decreases in importance, and as the social institution of work wanes as a principal social value.

A collapse of the work ethic may create despair among people who have been socialised to think of it as a central if not the most important value in their lives. Alternatively, social workers may facilitate ways of helping people connect within

their communities undertaking projects addressing health promotion, creative engagement, food production, and the development of local infrastructure, especially green, sustainable, and low-cost alternatives. In the hypermodern locality development, particularly in enriching participatory values, may be one of the most important shifts in moving from the modern to the hypermodern. Locality development can create new forms of employment as displaced workers come to appreciate the social impact they can make in their local communities.

Social work may respond to the hypermodern through public work projects akin to the Works Progress Administration in Depression era United States. Time for craft, the arts, performance, local construction, and personal development may rekindle the importance of local community centres, which under public auspices or the nonprofit sector may revitalize the tried and true in American social group work moving it from a clinical focus to its reinvention of democratic involvement. Social workers may collaborate with artists, information scientists, engineers, and health care personnel who together engage in community projects they design and implement to improve the environmental aesthetics of whole communities and even regions.

Such projects may result from government support or others may occur at local levels in which people collaborate in creating their own structures of mutual support, such as what occurs within the non-governmental or nonprofit sectors. Ultimately, however, funding of these alternatives may come from government at various levels of society, corporations that must contribute some of their profits to sustain their legitimacy, or a novel policy system that supports equity in work and income. As Brest and Born (2013) emphasize, corporations may need to prioritize social impact as a second rail to the maximization of profits alone. Funding will likely come from corporations and governments that are seeking societal stabilisation in the face of potentially massive labour dislocation, and the turmoil dislocation can create. Innovations in social entrepreneurship can occur through the consilience of microfinance and AI as tools for advancing innovation at local community levels (Allen, 2020).

Social work has numerous tools to engage in such practice; therefore, the ideas I offer are not utopian. The profession's innovations in group life, like the membership model, clubhouses, community centres, and mutual support can help people find meaning, address boredom, invest their surplus time, cognition, and energy in personal and community development through new careers they embrace in the social domain. Surplus time may become a reality of productive engagement in society as work becomes less salient in the lives of people whose economic security requires massive income support. A three-day work week may support the broad-based inclusion of people in the world of work and the world of their households and communities. Intergroup conflict or interclass conflict may fill the vacuum if government and corporate structures fail to invest in good wages, innovations in the fulfillment of surplus time, and social harmony.

The hypermodern may foster social activism when people realize that societal transition to a smart world will require if not demand worker protection and regulation of unbridled and unprecedented corporate profits (Koff, 2017). Deepening

public understanding of AI and its dark and bright sides can position social work for engaging people in public dialogue about the human use of smart technologies. The policy aim of activism here may translate not only into advocacy for universal income as a benefit but also advocacy for adequate household social income, the augmentation of which can be achieved through universal health care, productive strategies that prevent early movement into the workforce, and investment in creative and productive community service infrastructure. A salient area of social investment can be in community revitalization and the development of ecological resilience in the face of unbridled and life-threatening climate change. Societal responses to climate change through the development of community resilience may offer a new labour-intensive growth sector of the near future. This is somewhat paradoxical since the climate impact of AI is highly negative (Crawford, 2021).

Social work's innovations in fostering successful transitions can help people negotiate the movement from a work-centred society to a society focusing on public and civic engagement. With the availability of universal income supporting social consumption, surplus time, novel career patterns, groups may find commonalities in advancing the quality of environment, the quality of day, and the quality of life within their communities and so intergroup relations could improve. Social workers can be pivotal in reducing the uncertainty inherent among groups during periods of profound social transition that could otherwise result in conflict.

Social work itself may need to shift its focus to work and income as widespread social activism struggles politically in achieving equity in fulfilling the full and escalating costs of daily life involving housing, energy use, health care, transportation, and food. Concurrent with the disruption brought on by advanced technologies, the technologies themselves may reduce the cost of goods and services, lower energy costs, lower the cost of energy, and reduce the cost of living through innovations in healthcare, housing and food production.

In this way, the hypermodern may create more affordable goods and services, especially as government (with the support of wealthy corporations) socialize investments in local communities. Here both overarching societal policy structures focusing on income security, and local opportunities for productive engagement, especially in advancing community resilience in the face of massive climate change, could facilitate a productive transition out of the hypermodern. A renaissance may emerge globally as the dark ages of the hypermodern recede perhaps sometime in the 2060s as the threats of climate change abate, the pressures of human population decrease, cheap energy becomes ubiquitous, and social harmony prevails among groups once divided by conflicting identities.

Conclusion

Certainly, wealth disparities will likely continue to plague all societies, but especially the United States. Politically, wealthy corporations that survive deflationary forces in society given reductions in demand, perhaps dramatically, may find it to their advantage to pay more taxes, and invest more of their wealth in the development of local communities and group life.

The reinvigoration of revenue sharing in which the federal government invests directly in local municipalities, rural areas, and metroplexes, can arm social workers with the resources they need to advance local civic life. Local social and cultural infrastructure could expand under such resource schemes. And innovation in public financing of civic engagement can produce new forms of practice knowledge within the profession contributing to the rejuvenation of community practice and social action within social work education.

Doubtfully, social work or other professions can forestall the growing momentum of smart technology in society. What social work can do is to search for new and novel ways of organizing, especially at both societal, and local levels, to address the disappearance of traditional work or labour as we have come to know it. Richards (2018), however, challenges the intrusion of smart technology into society. He offers a theory in which technology itself amplifies what he calls the "human advantage," and identifies human virtues as the distinctive human qualities that will make human beings resilient in the age of the smart machine. Of the five virtues he flags is what he calls "creative freedom," the virtue that enables humans to create value for other humans, building on the other four virtues involving courage, antifragility, altruism, and collaboration.

These values hold evolutionary significance for the human species, and it is these virtues that social work practices. Valiant (2024) asserts that the locus of human civilisation is found in the educability of the human species. He underscores the quality of educability as making humans unique in the age of the smart machine. This too is consistent with the social work of the future—a profession devoted to advancing human capabilities with AI as an ally. Indeed, the progression of the hypermodern may favour the extension of education across the life course melding traditional methods with more experiential ones unfolding in communities, work settings, and new organizational structures (Dede & Richards, 2020).

It is here, in the human work of advancing human capabilities, that social work will find its most distinctive quality—helping humans connect with humans, the essential capability that can foster the evolution of human beings in the age of the smart machine (Pugh, 2024). Here the theme of intergroup collaboration will prove an important capability in making a fairer, more inclusive, and more equitable world (Khan, 2024). As the hypermodern recedes, a new global enlightenment may ensue in which human beings remain pivotal forces shaping a better world community.

References

Allen, D. W. (2020). Microfinance, entrepreneurship, and artificial intelligence: Prospects for poverty alleviation. *Development and Change*, *51*(2), 421–434. https://doi.org/10.1002/jsc.2356

Ananthaswamy, A. (2025). *Why machines learn: The elegant math behind modern AI*. Dutton.

Bean, M. (2024). *The skill code: How to save human ability in an age of intelligent machines*. Harper.

Bender, E., & Hanna, A. (2025). *The AI con: How to gith big tech's hype and create the future we want*. Harper.

Bloomberg Law AI. (2025a). *Alert: OpenAI, Anthropic, xAI hit with copyright lawsuits from writers*. December 22, 2025.

Bloomberg Law AI. (2025b). *Calls grow for AI incident database*. December 23, 2025.

Boyle, J. (2024). *The line: AI and the future of personhood*. MIT Press.

Brest, P., & Born, K. (2013). When can impact investing create real impact? *Stanford Social Innovation Review*, *11*(4), 22–31. https://doi.org/10.48558//JFWQ-GZ70.

Brockman, J. (2019). (Ed). *Possible minds: 25 ways of looking at AI*. Penguin.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Dede, C. J., & Richards, J. (2020). *The 60-year curriculum: New models for lifelong learning in the digital economy*. Routledge.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (AI Act)*. COM(2021) 206.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482–5

Flynn, J. R. (2000). *How to defend humane ideals*. University of Nebraska Press.

Fukuyama, F. (2018). *Identity: The demand for dignity and the politics of resentment*. Farrar, Straus, and Giroux.

Gabriel, M. (2022). *Moral progress in dark times*. Polity.

Harding, V. (2024). *AI needs you: How we can change AI's future and save our own*. Princeton University Press.

Ignatius, D. (2016). The brave new world of robots and lost jobs. *Washington Post*, August 16, 2016. Retrieved September 1, 2025. https://www.washingtonpost.com/opinions/the-brave-new-world-of-robots-and-lost-jobs/2016/08/11/e66a4914-5fff-11e6-af8e-54aa2e849447_story.html?utm_term=.83463e91928b.

Kateb, G. (2011). *Human dignity*. Belknap Press.

Khan, S. (2024). *Brave new words: How AI will revolutionize education*. Viking.

Kingsnorth, P. (2025). *Against the machine: One the unmaking of humanity*. Thesis.

Koff, S. (2017, March 3). Senator Sherrod Brown speaks at OSU, pushing for worker protections with loads of data. *Cleveland.Com*. https://www.cleveland.com/metro/2017/03/is_it_time_to_rethink_how_work.html

Matei, A. (2025). Systems are crumbling—But daily life continues. The dissonance is real. *The Guardian*. Retrieved May 22, 2025. https://www.theguardian.com/wellness/ng-interactive/2025/may/22/hypernormalisation-dysfunction-status-quo.

Meyerson, H. (2014). Technology and trade policy is pointing America toward a job apocalypse. *Washington Post*, March 26, 2014. Retrieved September 5, 2017. https://www.washingtonpost.com/opinions/harold-meyerson-technology-and-trade-policy-is-pointing-america-toward-a-job-apocalypse/2014/03/26/ba331784-b513-11e3-8cb6-284052554d74_story.html?utm_term=.dc0b38cee0c3

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, *33*(4), 659–684.

Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Penguin.

Moxley, D. (2025). *Responsive community building in high need contexts: The good organisation and the aesthetics of human development*. Palgrave Macmillan.

Nichols, T. (2024). *The death of expertise: The campaign against established knowledge and why it matters*. Oxford University Press.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Pasquinelli, M. (2023). *The eye of the master: A social history of Artificial Intelligence*. Verso.

Pepi, M. (2025a). *Against platforms: Surviving digital utopia*. Melville House.

Pepi, M. (2025b). The slop tax, a brief introduction: A cultural reinvestment framework for the AI age. Retrieved December 20, 2025. https://mikepepi.substack.com/p/the-slop-tax-a-brief-introduction.

Politico Toplines. (2025). *ICE's surveillance spending spree*. December 27, 2025.

Pugh, A. (2024). *The last human job: The work of connecting in a disconnected world*. Princeton University Press.

Rampell, C. (2015). The robots aren't threatening your job. *Washington Post*, April 9. Retrieved September 5, 2017. https://www.washingtonpost.com/opinions/dont-fear-the-robots/2015/04/09/e7ea1316-def3-11e4-a1b8-2ed88bc190d2_story.html?utm_term=.40a7fe1fecce

Richards, J. W. (2018). *The human advantage: The future of American work in an age of smart machines*. Crown Forum.

Schon, D. A. (1967). *Technology and change*. Dell Publishing.

Shirky, C. (2010). *Cognitive surplus: Creativity and generosity in a connected age*. Penguin.

Smith, G., O'Connor, C., & Zick, T. (2020). Tech worker resistance and grassroots innovation for social good. *New Media & Society*, *22*(7), 1165–1182.

Sunstein, C. R. (2004). *The second bill of rights: FDR's unfinished revolution and why we need it more than ever*. Basic Books.

Susskind, R., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press.

Togelius, J. (2024). *Artificial general intelligence*. MIT Press.

Valiant, L. (2024). *The importance of being educable: A new theory of human uniqueness*. Princeton University Press.

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A, Langhans, S.D., Tegmark, M., & Nerini, F.F. (2020). The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-019-14108-y

Walther, C. (2020). *Development, humanitarian action, and social welfare: Building bridges in the immediate aftermath of conflict*. Palgrave Macmillan.

Walther, C. (2021). *Adaptive leadership in development*. Palgrave Macmillan.

Wiener, N. (1950). *The human use of human beings: Cybernetics and society*. Da Capo Press.

Wilson, W. J. (1997). *When work disappears: The world of the new urban poor*. Vintage.

Winthrop, R., & McGivney, E. (2016). *Skills for a changing world: Advancing quality learning for vibrant societies*. Brookings.

Yale University. (2025). Socially assistive robotics: An NSF expedition in computing. Retrieved December 15, 2025. https://robotshelpingkids.yale.edu/overview.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. Basic Books.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Public Affairs.

4 Can Artificial Intelligence Be Controlled? A Social Science Perspective

Hermann Astleitner

A Problem-Sensitive Perspective on Artificial Intelligence

Some consider artificial intelligence (AI) as being our "last invention" or the "end of the human era", mainly because it is superior to humans, will replace them, and humans cannot control it (e.g., Barrat, 2023; Federspiel et al., 2023). Others see AI as epochally dangerous: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Center for AI Safety, n.d.). AI is

> a particular computer system or machine that has some of the qualities that the human brain has, such as the ability to interpret and produce language in a way that seems human, recognize or create images, solve problems, and learn from data supplied to it.
>
> (Cambridge Dictionary, n.d.)

AI systems and machines, and their use, are rapidly evolving, and the effects and consequences of AI technology cannot yet be accurately predicted. Currently, there are several different development options, such as artificial narrow intelligence (i.e., weak AI, designed to excel at a single specific task), artificial general intelligence (i.e., AI with human-level intelligence that works across multiple tasks and domains), and artificial superintelligence (i.e., AI with exceptional thinking skills that exceed human intelligence) (Kim et al., 2024). Here, I do not differentiate between these options but generally assume that as AI develops, the effectiveness of control mechanisms must improve accordingly. Some see AI as an important tool for solving the problems of humankind, such as those in medicine, education, economy, and agriculture (e.g., Astleitner, 2024; Qin et al., 2024). In this chapter, I do not focus on the promising potential benefits of AI but instead offer a problem-sensitive approach that emphasizes taking a preventative or interventionist perspective, especially in the event of AI-related challenges or disasters.

DOI: 10.4324/9781003627920-4
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

AI has some potential shortcomings, problems, or unwanted effects, which highlight the need for this critical forward-looking perspective:

- Bias: AI is not perfect and makes mistakes when processing data (e.g., Challen et al., 2019).
- Hallucinations: AI invents things that do not exist (e.g., Salvagno et al., 2023).
- Misuse: AI can be used in cheating, crimes, or wars to help achieve unethical goals (e.g., Johnson, 2019).
- Poisoning: AI training results can be manipulated when attackers add malicious examples to the training data (e.g., Wang et al., 2022).
- Surveillance: AI can infringe on privacy and be used as a tool for controlling people, to reinforce authoritarian structures in economics or politics (e.g., Unver, 2018).
- Replacement: AI can replace human labour, leading to unemployment and social unrest (e.g., Mutascu, 2021).
- Reciprocal relationship with human intelligence: AI can be used to replace human skills (e.g., thinking), leading to people not using or forgetting these skills, which could ultimately reduce human intelligence (e.g., Da Silveira & Lopes, 2023).
- Mutation: AI can self-improve and self-modify, but it can also mutate and become something that people do not want (e.g., Turchin, 2018).
- Uncontrollability: AI works in such a complex and untransparent manner that humans struggle to understand it and can no longer control it (e.g., Winkel, 2024).

This comprehensive list of problems related to AI underscores the importance of addressing how to control this potentially risky and overpowering technology. Controlling AI means regulating all the components of an AI computer system, including activities such as planning, implementing, evaluating, modifying, repairing, or shutting the system on and off (e.g., Soares et al., 2015). Controlling something means comparing a product or process with a reference value. When deviations from the reference value (or standard) are found, the system is changed so that the reference standard is likely to be achieved (e.g., Carver, 2004). That would also be the goal if one were to design a controlled AI.

Focus and Method: Analogical Reasoning between Evidence on Humans and Possible Behaviour of Artificial Intelligence

At present, it seems that the problems and potential negative effects of AI have been acknowledged, as multiple approaches to controlling AI have been proposed and developed (e.g., Tegmark, 2017). In this chapter, I present an overview of such approaches and examine them in the light of empirical evidence. Robust empirical evidence for controlling AI does not currently exist, or only fragments exist in computer science and related fields of research (e.g., Russell, 2019). I therefore

consider studies from the fields of psychology and education that address human intelligence and behaviour. In doing this, I draw an analogy between human intelligence and AI, similar to what other researchers have done in the past (e.g., Korteling et al., 2021).

I am aware that the "evidence problem" regarding humans does not necessarily apply to machines or computer software. However, if we want AI to approximate human intelligence, then this comparison is appropriate, especially in the absence of better evidence. Nevertheless, my work is not about providing an empirical assessment of the possibilities for controlling AI, but rather, I am trying to further develop and deepen a scientific perspective on the evidence problem using analogical reasoning (Ketokivi et al., 2017). Therefore, my work can be regarded as a social science-based conceptual analysis rather than an empirical test or evaluation related to computer science, mathematics, or statistics. From a methodological perspective, I combine the arguments on how AI can be controlled with insights from social science research on human intelligence and behaviour. Furthermore, I do not conduct a systematic review of empirical evidence but offer a more selective narrative review for triggering and supporting more problem-sensitive analyses (Pae, 2015).

Ways of Controlling Artificial Intelligence and Evidence from the Human Intelligence and Behaviour Fields

Hoping for Limits of Computation and Superintelligence

AI is computer-bound. Humans have had personal computers for decades, and they have had a major influence on our lives. But computer technologies have limitations due to a long list of characteristics that are unique to humans and not yet effectively replicated in AI, such as curiosity, imagination, intuition, emotions, passion, desires, pleasure, aesthetics, joy, purpose, wisdom, and humour (Braga & Logan, 2017). There have also been societal approaches to keeping computer technologies and their applications out of our lives as much as possible, such as digital detox (e.g., Radtke et al., 2022). From a human perspective, therefore, people might hope that AI remains constrained—that is, AI will not achieve superhuman intelligence and become too powerful. However, hope is associated with a passive belief in one's own control and often prevents people from acting (e.g., Miceli & Castelfranchi, 2010).

When comparing these hopes about the future of AI with social research on humans, we could look to the general development of human intelligence. Research in this area has examined whether people have become more intelligent through evolution, especially over recent decades. Such questions are being addressed in research on the Flynn Effect, which indicates an increase in average human intelligence over time in certain societies. However, more and more studies have found a negative Flynn Effect in many countries (e.g., Dutton et al., 2016). These studies are not always without controversy, but their results indicate it is unlikely that

as humans, we will develop superintelligence in the foreseeable future, especially without aids such as technical brain implants or performance-enhancing drugs.

Another human hope regarding AI superintelligence comes from social research. The hopes for the limits on computer technology and superintelligence (i.e., AI with intelligence surpassing that of humans) are partially shared by AI experts. The results of a survey by Müller and Bostrom (2016) show that AI experts assume that high-level machine intelligence should be achieved by around 2080, and by 30 years later (around 2110), machines should have superintelligence. These authors also report that the opinions of AI experts differ about the influence of superintelligence: about 52% think it will have a good impact on humanity, 17% are more or less neutral, and 31% think it will have a bad impact. In summary, and given the analogy with human intelligence, hopes about the limits of AI as a passive approach to controlling AI seem justified. It is worth noting that some AI experts and computer scientists consider it unlikely that superintelligence cannot be achieved by AI in the distant future. However, others believe that superintelligence is theoretically, and practically impossible (e.g., Alfonseca et al., 2021).

Self-Awareness and Self-Improvement

A more active approach to controlling AI is for AI systems to develop their own philanthropic consciousness and optimize themselves by learning from experience. As with humans, a higher level of AI self-awareness should come with the AI demonstrating a higher level of self-improvement and control over its own activities. However, this approach needs "strong AI" or "artificial general intelligence" based on, for example, advanced, and viable concepts of consciousness that can handle the mind–brain problem, or full AI autonomy, all of which is still scientifically and technically underdeveloped (e.g., Ng & Leung, 2020).

In the human context, such an approach raises the question of how people develop and what role intelligence plays in this. In this respect, there is evidence that crystallized (culturally knowledge-based) intelligence remains more or less stable, but fluid intelligence (as basic information processing), after increasing until early adulthood, declines over the course of a lifetime (Lindenberger, 2001). There is also an inverted U-shaped relationship between age and other intelligence-related developmental dimensions—such as wisdom, emotional regulation, and openness—which peaks between 40 and 60 years of age (Webster et al., 2012). The human perspective would therefore suggest that many cognitive and non-cognitive abilities decline as we get older. If what applies to humans also applies to AI, then the development of self-awareness and self-improvement in AI is constrained to a limited lifespan. However, in humans—AI analogous reasoning for limited self-improvement lies in age—which has not yet been examined in AI. Hence, we do not know if AI can age. However, if we assume that the data AI learns from is outdated, we can also assume that AI's ability to self-improve is limited. Hence, self-improvement is not a viable means of controlling AI, especially if it is not provided with the latest knowledge.

Being More Human: Artificial Emotional Intelligence

Making AI more human-like to control it involves equipping it with emotional skills—specifically, an emotion module that represents its emotional state, which it could use when processing information and making decisions (e.g., Cominelli et al., 2018). However, at the time of writing, an AI that can map, use, or experience a similar emotional range to humans is still in the development stage (Vistorte et al., 2024). Another problem is that artificial emotions might strongly differ in their goals compared to human emotions. It could be that emotional components will make AI more empathetic, potentially preventing harm to humans. However, human emotion can also lead to aggressive or unreasonable decisions, leading to uncontrollable behaviour that can hurt people, and this may also be the case for AI systems.

From a human perspective, the concept of emotional intelligence sits on the interface between emotions and intelligence. Emotional intelligence is about perceiving, understanding, regulating, and expressing emotions so they positively influence personal well-being and relationship building (e.g., Bastian et al., 2005). Research shows that emotional intelligence has positive effects on a wide range of experiences in human life, but it is difficult to learn, and its effects are often weak to moderate (e.g., Clarke, 2010). In addition, when emotional intelligence is combined with negative (dark) personality traits, it can have negative effects (Davis & Nichols, 2016). From a human perspective, therefore, integrating emotions into AI can have both positive and negative effects. Presently, we have no idea what will happen when different emotions are activated simultaneously in AI. There could be complex interactive processes and multiple unforeseen side-effects, similar to emotional outbursts or the reverse effects seen in polypharmacy (Sorter et al., 2022).

Laws of Robotics

Diverse versions of the laws of robotics (e.g., a robot may not injure a human being) or AI laws (e.g., the EU AI act) should guarantee that AI can be controlled through lawful regulation (e.g., McCauley, 2007). However, critics of such legal solutions argue that these laws fundamentally do not work from a purely legal perspective because they raise complex ethical problems, some of which are themselves unethical, or impose social constraints (e.g., Stokes, 2018).

Given my focus, the question arises as to whether human intelligence correlates with compliance with legal norms. If greater human intelligence means greater compliance with laws, then this could also mean that high-intelligence human-like AI also abides by laws. Research on human intelligence reveals several key insights in this respect. First, a high level of intelligence appears to protect against behaviour that deviates from the law (e.g., Ttofi et al., 2016). However, there is also the highly intelligent criminal mastermind, in which high intelligence contributes to the exact opposite of lawful behaviour (e.g., Oleson, 2016). From a social research perspective, it seems highly likely that intelligence only has an indirect influence on unlawful behaviour and that contextual conditions (e.g., the influence of a peer group) are more important (e.g., McGloin et al., 2004). When transferring the research context from humans to AI, the findings do not allow for a clear prediction

about how AI might behave if oriented toward laws in the future. In principle, a highly intelligent AI should strictly adhere to the law; however, this may be limited by contextual conditions. Laws are not independent; rather, it is clear that legal foundations are always anchored in conservative or progressive political views. For example, if we look at the current war in Gaza, where Israel has been using AI for war purposes, we can conclude that an important law of robotics, namely that machines must not harm people, has already been broken (The Economist, 2024). Thus, it seems that AI will harm people if other people or computer systems want it to and think it is right. Legal compliance may only become relevant once AI is fully independent of non-legal human entities or computer systems, which is not currently foreseeable.

Implementing Ethics

Like the laws of robotics, the implementing ethics approach to controlling AI assumes that a framework of ethical principles (about what is right and wrong) can be built for all AI activities and that AI should become a beneficial and moral machine. From this perspective, AI should only promote or create things that are useful to humans. It should also learn to evaluate situations and decisions according to moral principles and act accordingly. According to such ethical principles, AI should have beneficence (e.g., promote well-being), nonmaleficence (e.g., protect privacy), autonomy (e.g., the power to make decisions), justice (e.g., avoid unfairness), and explicability (e.g., enable principles through intelligibility and accountability) (Floridi, 2023). There are several difficulties here, however. For example, it is difficult to reach agreement on ethical rules because they are often influenced by political views or economic interests. It must also be assumed that different ethical rules apply to different areas of life or fields of research, but there is a long way from applying these rules to their acceptance and effective implementation in everyday contexts (for a critique of AI ethics, see Cole et al., 2022).

Research on human intelligence has addressed the question of whether humans can learn ethical rules and apply them consistently in everyday life. The findings show that human beings are able to learn ethical rules and the necessary training measures have an average impact (e.g., Mulhearn et al., 2017). There is also evidence indicating that higher intelligence is related to a higher emotional sensitivity, a greater concern for others, greater self-identification as a moral person, and more prosocial behaviour (Guo et al., 2019). However, whether people accept or break rules is not related to their level of intelligence (e.g., Dubois et al., 2015). Research also shows that there can be a collective violation of rules, or people exhibit unethical behaviour for the benefit of a group (e.g., Krause et al., 2021). In addition, the implementation and application of rules depend on many different and interacting factors (e.g., leadership or implementation aids) (e.g., Kelly et al., 2021). The same applies here as with the laws of robotics: whether ethical principles are consistently enforced in the use of AI depends on a variety of mutually exclusive factors that are difficult to coordinate and even more difficult to control.

Transparency, Openness, and Explainability

Another approach to controlling AI is ensuring that the models, algorithms, data, software structures, etc., that underlie an AI system are transparent, openly accessible, and explainable (e.g., Larsson & Heintz, 2020). The assumption behind this is that if everything about AI is clear and understandable, then we have a good basis for controlling it. However, evidence from AI research suggests that often, for example, the data used for training and testing AI algorithms lacks transparency, which makes it difficult to assess the robustness and potential biases of AI systems (Daneshjou et al., 2021).

From a human perspective, and in a meritocracy and competitive economy, a lack of transparency often represents a competitive advantage (e.g., Crittenden et al., 2015). We also have intransparency-related deception and lying as an everyday phenomenon, as well as dark personalities with a strong tendency to hide knowledge (e.g., Denault et al., 2022). Despite decades of psychological research and significant progress, human behaviour is neither adequately explained nor sufficiently predictable, controllable, or changeable (Hagger et al., 2020). Overall, therefore, one thing seems clear regarding the human–AI analogy: AI is an economic factor and if intransparency or hiding knowledge are competitive advantages, it is unlikely that transparency will work as an effective AI control factor.

Risk and Safety Management

Anyone who works with AI has to make decisions under conditions of uncertainty (Hermstrüwer, 2020). Taking this uncertainty into account, AI and risk and safety management can be combined by focusing on concepts such as the precautionary principle (for protecting people from harm) or governance (including legal frameworks for monitoring and managing) (e.g., Taeihagh, 2021). Risk and security management in AI is based on benchmarks such as fairness, reliability, honesty, or non-malicious use, but these are still insufficiently implemented (Ren et al., 2024).

When comparing such approaches with human risk and security or disaster management, it is obvious that this is a complex task that requires a high level of coordination between different actors and multiple activities, including hazard assessment (exposure analysis, forecast, vulnerability assessment, resource assessment), risk management (risk communication, identification, evaluation, treatment, monitoring), and management actions (prevention/warning, mitigation, response, recovery, learning development) (Nojavan et al., 2018). When reflecting on such activities within the AI context, it becomes apparent that AI is operated independently by different developers using different tools in different countries and jurisdictions, under different political and economic spheres of influence. Coordinated risk and security management in the event of an AI disaster therefore seems unlikely.

Monitoring and Evaluation

From the discussion so far, AI is undoubtedly fraught with bias. Multiple types of bias—such as systemic bias (e.g., institutional bias), human bias (e.g., groupthink),

and statistical/computational bias (e.g., errors in data detection)—can be identified when confronting AI with standards for monitoring and evaluation (Schwartz et al., 2022). There are performance tests for evaluating AI, but these must be developed for different areas of application and enable a diverse assessment of AI performance (e.g., Naser & Alavi, 2023).

When applied to human learning, different forms of monitoring and evaluation (e.g., formative assessments) have a positive and significant effect on many fields of learning and performance (e.g., Lee et al., 2020). What matters is both the quality of the feedback in terms of opportunities for improvement, but also how this feedback is processed with respect to goals, meaning-making, and decisions (Lui & Andrade, 2022). Monitoring and evaluation could prove to be highly effective control tools, provided that this process is monitored by a superior or external institution and made the subject of quality development measures. However, the main problem is that monitoring, evaluation, and quality assurance only work effectively when recognized standards are in place, and if all services or service deficits are fully accessible. This may be at least partially true for companies and educational institutions, but not yet for the field of AI. Ultimately, implementing this approach to controlling AI means that many of the other above-mentioned approaches to controlling it need to have already been successfully implemented. But it is doubtful this will happen any time soon. However, it might be possible to use control mechanisms from human areas that could ensure high levels of control. Examples of areas with high control could include nuclear power applications or aviation. For example, like flight control systems AI systems should be based on layered safeguards including permanent monitoring related to irregular activities, unauthorized systems interactions, or mechanisms to reverse harmful manipulations (Duranton & Lyon, 2025).

Discussion

This study on controlling AI provides no reassurance about its risks; our analogous view suggests that it will be very difficult to control AI, at least in the immediate future. The results of my conceptional analysis based on analogical reasoning are partly consistent with, for example, findings from similar conceptual work on controlling autonomous AI weapons systems and evidence on the uncontrollability of AI from AI safety research (Yampolskiy, 2024). However, future research can proceed calmly, but with greater awareness, recognizing technological revolution is essential, and has changed the world significantly—as in the case of social media (e.g., Astleitner & Schlick, 2025). While AI technologies are still in their infancy, we perhaps need to take a nurturing and responsible approach—that is guided by rules and education, much like we do with children—as an effective framework for future research and development (e.g., Gopnik, 2017). Authoritative "AI parents" should be "demanding and responsive, controlling but not restrictive" (Klein & Ballantine, 2001, p. 46). While such a view could reveal our naivety about AI, it could also increase our control capabilities.

Of course, my exploration of controlling AI from a human perspective is speculative. I used the analogy between human behaviour and AI to explore the cross-connections based on empirical studies from social research. I selected several important approaches and found they all have significant shortcomings. An important outcome of this work is the conclusion that to be effective in the future, approaches to controlling AI need to be further optimized and researched. Future research on controlling AI should primarily aim at security, which includes human, national, and existential security, as well as human trust in AI (Glikson & Woolley, 2020). Security and trust are integrative concepts that might allow the merging of different approaches to controlling AI. Based on my findings, it seems certain that the existing approaches to controlling AI should be used in a combined and coordinated manner, as each individual control option could be ineffective. Monitoring and evaluation with a quality such as that achieved in aviation appear to offer the most promising and effective approach and could foster the integration of other approaches. However, implementing the control options must be coupled with a reflective intervention program that uses sophisticated knowledge about behaviour change (e.g., Hagger et al., 2020). Computer scientists need to possess this kind of knowledge, or else cooperate with researchers from the field of psychology or education.

Finally, in my study, I may have overlooked other important control options. Perhaps the strongest controlling factor for AI will be its economic potential. If profits are made with AI, then it will be expanded further. If these profits fail to materialize, as has largely been the case in the past, fewer resources will flow into AI development. However, an "automated neoliberalism" (based on a technology-optimized and accelerated capitalism as well as AI monopolies or oligopolies) is more likely to emerge here than an economy based on an AI tool built on wisdom and love (e.g., Birch, 2020). AI serves commercial purposes that are difficult to control. At the very least, with every request or prompt, one should ask the AI who exactly it is working for and what economic and other goals the operator of the AI is pursuing (Reed, 2025). There is another problem related to economic limits that could prove to be an indirect control mechanism in the future: AI could run out of training data. Restrictions on data use, limited in-depth learning capacity, or poor quality training data could hinder the further development of AI (e.g., Hagendorff, 2021).

Nyholm (2023) presents another important argument in the AI control debate that we have not considered here: the more similar AI becomes to humans, the greater the likelihood that we will not want to exercise full control over it. Given these complex circumstances, as argued by Salmi (2023), it seems clear that, in the long run, controlling AI must be ensured throughout a comprehensive, evidence-based, and fully accountable democratic process that balances diverse interests while weighing up the advantages and disadvantages of AI. Until this is actually implemented, at least the rule should apply that no human being can be forced to follow instructions of AI. AI may not harm humans on its own, but history teaches us that humans always use the latest technologies to satisfy their needs and desires, which sometimes means harming others. Until we are sure, we should take the potential of AI as seriously as the desire for a better life, but also the greed for money and power.

Acknowledgment

I extend my thanks to Hannah Zischg (Paris Lodron University of Salzburg) for her assistance in the literature search and her practical attempts to explore the limits of AI.

References

Alfonseca, M., Cebrian, M., Anta, A. F., Coviello, L., Abeliuk, A., & Rahwan, I. (2021). Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research*, *70*, 65–76. https://doi.org/10.1613/jair.1.12202

Astleitner, H. (2024). We have big data, but do we need big theory? Review-based remarks on an emerging problem in the social sciences. *Philosophy of the Social Sciences*, *54*(1), 69–92. https://doi.org/10.1177/00483931231188825

Astleitner, H., & Schlick, S. (2025). The social media use of college students: Exploring identity development, learning support, and parallel use. *Active Learning in Higher Education, 26*(1), 231–254. https://doi.org/10.1177/14697874241233605

Barrat, J. (2023). *Our final invention. Artificial intelligence and the end of the human era*. Quercus.

Bastian, V. A., Burns, N. R., & Nettelbeck, T. (2005). Emotional intelligence predicts life skills, but not as well as personality and cognitive abilities. *Personality and Individual Differences*, *39*(6), 1135–1145. https://doi.org/10.1016/j.paid.2005.04.006

Birch, K. (2020). Automated neoliberalism? The digital organisation of markets in technoscientific capitalism. *New Formations*, *100*(100–101), 10–27. https://doi.org/10.3898/NewF:100-101.02.2020

Braga, A., & Logan, R. K. (2017). The emperor of strong AI has no clothes: Limits to artificial intelligence. *Information*, *8*(4), 156. https://doi.org/10.3390/info8040156

Cambridge Dictionary. (n.d.). Artificial intelligence. *In Cambridge dictionary*. Retrieved August 26, 2024. https://dictionary.cambridge.org/dictionary/english/artificial-intelligence

Carver, C. S. (2004). Self-regulation of action and affect. In R. F. Baumeister & K. D. Vohs (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 13–39). Guilford Press.

Center for AI Safety (n.d.). *Statement on AI risk*. aistatement.com

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*(3), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

Clarke, N. (2010). Developing emotional intelligence abilities through team-based learning. *Human Resource Development Quarterly*, *21*(2), 119–138. https://doi.org/10.1002/hrdq.20036

Cole, M., Cant, C., Ustek Spilda, F., & Graham, M. (2022). Politics by automatic means? A critique of artificial intelligence ethics at work. *Frontiers in Artificial Intelligence*, *5*, 869114. https://doi.org/10.3389/frai.2022.869114

Cominelli, L., Mazzei, D., & De Rossi, D. E. (2018). SEAI: Social emotional artificial intelligence based on Damasio's theory of mind. *Frontiers in Robotics and AI*, *5*, 6. https://doi.org/10.3389/frobt.2018.00006

Crittenden, W. F., Crittenden, V. L., & Pierpont, A. (2015). Trade secrets: Managerial guidance for competitive advantage. *Business Horizons*, *58*(6), 607–613. https://doi.org/10.1016/j.bushor.2015.06.004

Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, *157*(11), 1362–1369. https://doi.org/10.1001/jamadermatol.2021.3129

Da Silveira, T. B. N., & Lopes, H. S. (2023). Intelligence across humans and machines: A joint perspective. *Frontiers in Psychology*, *14*, Article 1209761. https://doi.org/10.3389/fpsyg.2023.1209761

Davis, S. K., & Nichols, R. (2016). Does emotional intelligence have a "dark" side? A review of the literature. *Frontiers in Psychology*, *7*, Article 1316. https://doi.org/10.3389/fpsyg.2016.01316

Denault, V., Talwar, V., Plusquellec, P., & Larivière, V. (2022). On deception and lying: An overview of over 100 years of social science research. *Applied Cognitive Psychology*, *36*(4), 805–819. https://doi.org/10.1002/acp.3971

Dubois, D., Rucker, D. D., & Galinsky, A. D. (2015). Social class, power, and selfishness: When and why upper and lower class individuals behave unethically. *Journal of Personality and Social Psychology*, *108*(3), 436–449. https://doi.org/10.1037/pspi0000008

Duranton, S., & Lyon, V. (2025, March 19). What AI can learn from aviation safety. *Project Syndicate*. https://www.project-syndicate.org/commentary/ai-industry-must-adopt-robust-safety-protocols-by-sylvain-duranton-and-vanessa-lyon-2025-03

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn Effect: A systematic literature review. *Intelligence*, *59*, 163–169. https://doi.org/10.1016/j.intell.2016.10.002

Federspiel, F., Mitchell, R., Asokan, A., Umana, C., & McCoy, D. (2023). Threats by artificial intelligence to human health and human existence. *BMJ Global Health*, *8*(5), Article e010435. https://doi.org/10.1136/bmjgh-2022-010435

Floridi, L. (2023). *The ethics of artificial intelligence. Principles, challenges, and opportunities*. Oxford University Press.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gopnik, A. (2017). An AI that knows the world like children do. *Scientific American Mind*, *28*(5), 21–28. https://www.jstor.org/stable/27102278

Guo, Q., Sun, P., Cai, M., Zhang, X., & Song, K. (2019). Why are smarter individuals more prosocial? A study on the mediating roles of empathy and moral identity. *Intelligence*, *75*, 1–8. https://doi.org/10.1016/j.intell.2019.02.006

Hagendorff, T. (2021). Linking human and machine behaviour: A new approach to evaluate training data quality for beneficial machine learning. *Minds and Machines*, *31*(4), 563–593. https://doi.org/10.1007/s11023-021-09573-8

Hagger, M. S., Cameron, L. D., Hamilton, K., Hankonen, N., & Lintunen, T. (2020). (Eds). *The handbook of behaviour change*. Cambridge University Press.

Hermstrüwer, Y. (2020). Artificial intelligence and administrative decisions under uncertainty. In T. Wischmeyer & T. Rademacher (Eds.), *Regulating artificial intelligence* (pp. 119–223). Springer.

Johnson, J. (2019). Artificial intelligence & future warfare: Implications for international security. *Defense & Security Analysis*, *35*(2), 147–169. https://doi.org/10.1080/14751798.2019.1600800

Kelly, E. M., Greeny, K., Rosenberg, N., & Schwartz, I. (2021). When rules are not enough: Developing principles to guide ethical conduct. *Behaviour Analysis in Practice*, *14*, 491–498. https://doi.org/10.1007/s40617-020-00515-x

Ketokivi, M., Mantere, S., & Cornelissen, J. (2017). Reasoning by analogy and the progress of theory. *Academy of Management Review*, *42*(4), 637–658. https://doi.org/10.5465/amr.2015.0322

Kim, H., Yi, X., Yao, J., Lian, J., Huang, M., Duan, S., Bak, J., & Xie, X. (2024). The road to artificial superintelligence: A comprehensive survey of superalignment. arXiv: 2412.16468. Retrieved January 31, 2025. https://doi.org/10.48550/arXiv.2412.16468

Klein, H. A., & Ballantine, J. (2001). For parents particularly: Raising competent kids: The authoritative parenting style. *Childhood Education*, *78*(1), 46–47. https://doi.org/10.1080/00094056.2001.10521689

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus artificial intelligence. *Frontiers in Artificial Intelligence*, *4*, Article 622364. https://doi.org/10.3389/frai.2021.622364

Krause, J., Romanczuk, P., Cracco, E., Arlidge, W., Nassauer, A., & Brass, M. (2021). Collective rule-breaking. *Trends in Cognitive Sciences*, *25*(12), 1082–1095. https://doi.org/10.1016/j.tics.2021.08.003

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2), 1–16. https://doi.org/10.14763/2020.2.1469

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in US K-12 education: A systematic review. *Applied Measurement in Education*, *33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Lindenberger, U. (2001). Lifespan theories of cognitive development. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioural sciences* (pp. 8848–8854). Elsevier.

Lui, A. M., & Andrade, H. L. (2022). The next black box of formative assessment: A model of the internal mechanisms of feedback processing. *Frontiers in Education*, 7, Article 751548. https://doi.org/10.3389/feduc.2022.751548

McCauley, L. (2007). AI Armageddon and the three laws of robotics. *Ethics and Information Technology*, *9*, 153–164. https://doi.org/10.1007/s10676-007-9138-2

McGloin, J. M., Pratt, T. C., & Maahs, J. (2004). Rethinking the IQ–delinquency relationship: A longitudinal analysis of multiple theoretical models. *Justice Quarterly*, *21*(3), 603–635. https://doi.org/10.1080/07418820400095921

Miceli, M., & Castelfranchi, C. (2010). Hope: The power of wish and possibility. *Theory & Psychology*, *20*(2), 251–276. https://doi.org/10.1177/0959354309354393

Mulhearn, T. J., Steele, L. M., Watts, L. L., Medeiros, K. E., Mumford, M. D., & Connelly, S. (2017). Review of instructional approaches in ethics education. *Science and Engineering Ethics*, *23*, 883–912. https://doi.org/10.1007/s11948-016-9803-0

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In V. C. Müller (Eds.), *Fundamental issues of artificial intelligence* (pp. 555–572). Springer.

Mutascu, M. (2021). Artificial intelligence and unemployment: New insights. *Economic Analysis and Policy*, *69*, 653–667. https://doi.org/10.1016/j.eap.2021.01.012

Naser, M. Z., & Alavi, A. H. (2023). Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. *Architecture, Structures and Construction*, *3*(4), 499–517. https://doi.org/10.1007/s44150-021-00015-8

Ng, G. W., & Leung, W. C. (2020). Strong artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, *7*(1), 63–72. https://doi.org/10.1142/S2705078520300042

Nojavan, M., Salehi, E., & Omidvar, B. (2018). Conceptual change of disaster management models: A thematic analysis. *Jàmbá: Journal of Disaster Risk Studies*, *10*(1), Article a451. https://doi.org/10.4102/jamba.v10i1.451

Nyholm, S. (2023). A new control problem? Humanoid robots, artificial intelligence, and the value of control. *AI and Ethics*, *3*(4), 1229–1239. https://doi.org/10.1007/s43681-022-00231-y

Oleson, J. C. (2016). *Criminal genius: A portrait of high-IQ offenders*. University of California Press.

Pae, C. U. (2015). Why systematic review rather than narrative review? *Psychiatry Investigation*, *12*(3), 417–419. https://doi.org/10.4306/pi.2015.12.3.417

Qin, Y., Xu, Z., Wang, X., & Skare, M. (2024). Artificial intelligence and economic development: An evolutionary investigation and systematic review. *Journal of the Knowledge Economy*, *15*(1), 1736–1770. https://doi.org/10.1007/s13132-023-01183-2

Radtke, T., Apel, T., Schenkel, K., Keller, J., & von Lindern, E. (2022). Digital detox: An effective solution in the smartphone era? A systematic literature review. *Mobile Media & Communication*, *10*(2), 190–215. https://doi.org/10.1177/20501579211028647

Reed, J. (2025, March 12th). Tim Berners-Lee wants to know: 'Who does AI work for?'. *CNET*. https://www.cnet.com/tech/services-and-software/tim-berners-lee-wants-to-know-who-does-ai-work-for/

Ren, R., Basart, S., Khoja, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Pan, A., Mukobi, G., Kim, R. H., Fitz, S., & Hendrycks, D. (2024). Safetywashing: Do AI safety benchmarks actually measure safety progress? https://doi.org/10.48550/arXiv.2407.21792

Russell, S. (2019). *Human compatible. AI and the problem of Control*. Penguin Books.

Salmi, J. (2023). A democratic way of controlling artificial general intelligence. *AI & Society*, *38*(4), 1785–1791. https://doi.org/10.1007/s00146-022-01426-x

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Artificial intelligence hallucinations. *Critical Care*, *27*(1), Article 180. https://doi.org/10.1186/s13054-023-04473-y

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence*. US Department of Commerce, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.1270

Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). Corrigibility. In T. Walsh (Ed.), *Artificial intelligence and ethics AAAI technical report WS-15-02* (pp. 74–82). AAAI Press. https://cdn.aaai.org/ocs/ws/ws0067/10124-45900-1-PB.pdf

Sorter, M., Chua, J., Lamy, M., Barzman, D., Ryes, L., & Shekhtman, J. A. (2022). Management of emotion dysregulation and outbursts in children and adolescents. *Current Psychiatry Reports*, *24*(3), 213–226. https://doi.org/10.1007/s11920-022-01325-4

Stokes, C. (2018). Why the three laws of robotics do not work. *International Journal of Research in Engineering and Innovation (IJREI)*, *2*(2), 121–126. https://ijrei.com/table/volume-2/issue-2

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Tegmark, M. (2017). *Life 3.0. Being human in the age of artificial intelligence*. Vintage Books.

The Economist. (2024, April 11). Israel's use of AI in Gaza is coming under closer scrutiny. https://www.economist.com/middle-east-and-africa/2024/04/11/israels-use-of-ai-in-gaza-is-coming-under-closer-scrutiny

Ttofi, M. M., Farrington, D. P., Piquero, A. R., Lösel, F., DeLisi, M., & Murray, J. (2016). Intelligence as a protective factor against offending: A meta-analytic review of prospective longitudinal studies. *Journal of Criminal Justice*, *45*, 4–18. https://doi.org/10.1016/j.jcrimjus.2016.02.003

Turchin, A. (2018). Levels of self-improvement in AI and their implications for AI safety. *PhilPapers*. https://philpapers.org/rec/TURLOS

Unver, A. (2018). Artificial intelligence, authoritarianism and the future of political systems. *EDAM Research Reports*. https://ssrn.com/abstract=3331635

Vistorte, A. O. R., Deroncele-Acosta, A., Ayala, J. L. M., Barrasa, A., López-Granero, C., & Martí-González, M. (2024). Integrating artificial intelligence to assess emotions in learning environments: A systematic literature review. *Frontiers in Psychology*, *15*, Article 1387089. https://doi.org/10.3389/fpsyg.2024.1387089

Wang, C., Chen, J., Yang, Y., Ma, X., & Liu, J. (2022). Poisoning attacks and countermeasures in intelligent networks: Status quo and prospects. *Digital Communications and Networks*, *8*(2), 225–234. https://doi.org/10.1016/j.dcan.2021.07.009

Webster, J. D., Westerhof, G. J., & Bohlmeijer, E. T. (2012). Wisdom and mental health across the lifespan. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *69*(2), 209–218. https://doi.org/10.1093/geronb/gbs121

Winkel, M. (2024). Controlling the uncontrollable: The public discourse on artificial intelligence between the positions of social and technological determinism. *AI & Society*, 1–13. https://doi.org/10.1007/s00146-024-01979-z

Yampolskiy, R. V. (2024). *AI unexplainable, unpredictable, uncontrollable*. CRC Press.

5 The Politics of Information and Communication Technologies (ICTs), Artificial Intelligence (AI), and Implications for Social Work, Justice, and Advocacy

Carolyn Noble

Introduction

There is no doubt that Information and Communication Technologies (ICTs) are infiltrating all aspects of everyday life. Imagine life without internet access, or a computer or mobile phone. Although various communication technologies have been with us for many decades, the introduction and growth of Artificial Intelligence (AI) programmers and tools, whose access is facilitated and supported by the growth and sophistication of ICTs, are almost unpreceded in their adoption, influence, and benefits. Computer systems capable of tasks once thought to require human intelligence have transformed communication, education, work, business, and social and political engagement. Where technology and access are available, no area of human activity is spared from interaction with ICTs and AI tools. Banking, government and health services, health reports, fines, text messages, online education, teaching and learning, and social assistance are accessed and communicated via these technologies. AI programmers and technologies are readily available for use in many important aspects of civil life such as employment, commerce, social media, education, the military, medicine, and engineering. In fact, we may be the last generation to experience a pre-digital reality. It is crucial, then, to discuss the development of ICTs and AI and their impact on the social fabric and social work, even as their impact is yet to be stabilised around a recognisable set of patterns (Bailey et al., 2022)

The Symbiotic Relationship between ICTs and AI

ICT refers to the wide range of tools and resources that are used to transmit, store, create, share, or exchange information. ICT encompasses physical hardware such as computers and smartphones, software including messaging applications, databases, and operating systems, as well as networks such as the internet and mobile communication systems.

Current AI technologies represent advanced computational capabilities and include machine learning (ML), deep learning (DL), computer vision, content creation platforms (such as Instagram, X, and Facebook), chatbots (like ChatGPT, Gemini, and Copilot), self-driving cars, and robotics, among others (HEPA, n.d., beDots, 2025).

DOI: 10.4324/9781003627920-5

This chapter has been made available under a CC-BY-NC-ND 4.0 license.

ICTs provide the foundational infrastructure—comprising hardware, networking systems, and large datasets—that is essential for training and operating various AI models and programmes. The development and operation of AI technologies depend on this environment. Both ICT and AI function in tandem: while ICT delivers the necessary environment for AI innovation and operation, AI tools, in turn, can automate processes, optimise network communications, and increase the efficiency of ICT systems.

ICT supports the development of AI by offering programming languages, software development tools, and platforms for deploying AI applications. Conversely, AI enhances ICT by enabling smarter automation, real-time data analysis, and improved decision-making. AI systems require large amounts of data to learn and evolve, and it is the role of ICT to collect, store, and process this data effectively (HEPA, n.d.; beDots, 2025).

Whereas ICT serves as the broader framework that covers all digital communication technologies, AI acts as the advanced capability that makes those systems more intelligent, efficient, and autonomous. The integration of ICT and AI has broadened the scope of potential applications, including data analysis, cybersecurity, smart devices, telemedicine, e-government systems, predictive policing, digital identity management, language translation tools, and generative AI for the arts. These AI-powered tools continuously shape new possibilities, and applications in areas such as education, healthcare, security, environmental protection, transportation, industry, science, research, and automation, providing significant benefits to everyday life (HEPA, n.d., beDots, 2025).

However, as ICT technologies and AI tools advance rapidly, social inequality and exclusion risk worsening for those on the periphery of society—individuals with limited or no access to communication technologies and supporting tools and the benefits that come with them. This creates a clear digital divide between users and non-users, deepening existing disparities between them.

The Digital Divide

Despite the seemingly pervasiveness of some types of communication technologies that purport to bring development benefits with their use and access, many nations, regions, groups, and individuals have been relatively or absolutely excluded from their benefits (van Dijk, 2020). As of 2023, around 2.6 billion people globally still do not have access to the internet**,** and billions more lack regular access to computers and smartphones according to the International Telecommunication Union (ITU). That is, one-third of the world's population does not have access to these technologies, either at work, school, and/or home, where age, income, ability, gender, isolation, political instability, and indigeneity are contributing factors. Without access to ICT, internet broadband and AI tools and technologies social and economic disparities, between the rich and poor, the Global North and Global South, rural and remote living, marginalised communities, and Indigenous peoples are in danger of becoming further entrenched. The disparities in access to ICT devices, infrastructure, and AI tools and programmes among peoples, countries, and regions have become another

fertile ground to breed and perpetuate social exclusion and exacerbate poverty. As so much of social interaction is online and available for those who have access to the requisite technologies and ICTs, those without this online interaction and adequate and accessible tech infrastructure are prevented from fully accessing their social, cultural, and political rights as citizens. It can be argued that these barriers are precursors to, or even act as causes of, the digital divide.

Factors Contributing to Digital Divide

There is no escape from the digital landscape, and the move to digitalisation continues to broaden inequalities, which involve the unequal distribution of resources, opportunities, and privileges across different social groups' (Cong & Quinn, 2025, 3). This lack of access creates a schism between the people using, obtaining, and affording digital services and those digitally excluded. This raises questions about who is developing new technologies, for whose benefits, who mostly consumes them and has full access, and who are the content creators (Yadav, Yadav & Pawar, 2025). Currently, user-generated content originates from a small segment (white, young, tech-savvy men and boys) of predominately western countries where cultural values, lived experiences, and interpretations of human rights (as an example) may be flattened or homogenised.

Key elements contributing to the digital divide include the actual infrastructure (fluctuating bandwidth and accessibility), geographical isolation, poverty, political instability, government censorship as well as affordability, quality of devices, digital literacy, inaccessible websites, financial hardship, lack of assistive technology, limited digital literacy support, and relevant content creation. While the divide is not monolithic, certain social consequences can be attributed to its impact across countries and regions where access is limited (Orellana-Rodriguez, 2024). Broadly it encompasses three levels: access, skills, and use and benefits. As internet use grows, the digital divide now includes not only access and skills but also qualitative factors—such as differences in user abilities and structural inequalities like gender, race, ethnicity, disability, income, and age—which affect both quality of access and participation in civic life (Orellana-Rodriguez, 2024; Heeks, 2022). As inequality and social exclusion are major challenges facing the world, there are significant concerns about the contribution of the digital divide to further this inequality and social exclusion, which cast a shadow over the benefits attributed to its development.

Social Exclusion

Social exclusion is understood as the breakdown of the structural, cultural, and moral ties which bind the individual to society. For example, in many rural areas, access to technology and infrastructure and the capacity and speed of broadband are crucial for meaningful connectivity, so when the quality, capacity, speed, and affordability are compromised, they present ongoing challenges. Older adults and people with disabilities face additional barriers in accessing and using technologies.

This is true for low-income individuals as more social services are moving online, where internet access, e-government schemes, data plans, skills and knowledge use, and communication platforms present significant barriers. Further, for some disadvantaged children without home broadband, learning starts and stops in the classroom, where computer availability is limited to school hours and teaching terms. This is also true for Indigenous populations who are marginalised both from access, resources, and representation in the development of AI technology, further exacerbating their poverty, oppression, social isolation, and invisibility.

People affected by the digital divide lack access to online health, education, economic, governance, and social benefits, and their data is often excluded from government initiatives that support digitally connected citizens. For those who are digitally connected and aware, it means that they are confident in appropriately accessing, managing, integrating, evaluating, analysing, synthesising, and creatively use a range of digital technologies for self-development, and civic, cultural, and social engagement in the digitalised society. They can, as a result, engage in constructive social interaction in their daily activities and social life and to critically reflect upon this process and the value it brings (Cong & Quinn, 2025). Given that much of today's communication and interaction takes place on various internet platforms and through various ICTs the risk of a dual and fragmented society—a division between the haves and the have nots is—ripe for heralding a new source of inequality (Heeks, 2022; Hickey & du Toit, 2013). Social isolation and marginalisation prevent those excluded from participating in political, economic, social, health, and cultural activities, resulting in what Molala and Makhubele call "the worst kind of social injustice" (p.194).

Global South, Digital Divide, and Economic/Commercial Exploitation

In the Global South, digital exclusion is also linked to the inequality of poverty and exploitation. People who live physically, socially, and economically on the margins and are excluded from formal employment and welfare are easily captured into digital systems that are economically, culturally, and technologically disadvantageous. This explains why different incorporations into the digital system in the Global South exacerbate historical inequity between different groups, even as access has increased (Heeks, 2022). For example, small businesses and farmers trading on and using digital platforms, workers employed in the tech sector, and communities that have been digitally mapped have experienced their financial and commercial value extracted for profit in the Global North platforms, leaving too little value at the site of extraction, creating an uneven power structure and tiered economy (Heeks, 2022; Steyaert & Guld, 2009). While some gig workers and entities in the Global South have some limited success (such as building freelance careers in tech and design) and using ICT and AI tools, many experience inequality between different groups that use the same ICTs and AI technologies, but whose position of power and control are unequal and exploitative. To illustrate, TripAdvisor's concentration of market power, control of information, and internet sites results in significant

levels of capital accumulation outside of the Global South, thus enabling a more advantaged group to extract disproportionate value from the work or resources of another, less advantaged group (i.e. those providing the actual services). The expropriation of traditional community knowledge (e.g. health medicine, cultural practices, spiritual beliefs, not to mention everyday life expertise and judgement) into privately owned digital systems is a further example of adverse digital incorporation and exploitative nature of power differences between the Global North and South's use and access to ICTs and AI-powered tools providing little space or opportunity for the Global South entities and workers to exert some control over their adverse digital incorporation (Arora et al., 2023; Heeks, 2022).

This skewed value exchange extends the technological and productivity gap between the Global North and South as well (Arora et al., 2023; Heeks, 2022). For example, gig workers in the Global South are largely invisible working to clean up and refine algorithm development for the benefit of those using algorithms in the Global North. While other gig workers are lured into the crime economy to alleviate poverty where they are being ensnared into complex criminal ecosystems that operate under the guise of legitimate employment. These systems often masquerade as call centres or tech support hubs but are in fact scam compounds—sites of forced labour, sexual coercion, cyber fraud, and arms, drug, and human trafficking (Heeks, 2022; Hickey & du Toit, 2013), thereby perpetuating harm and obscuring the lived realities of those employed in these platforms. Additionally, as advantaged groups in the North have control of the design of the system, there is asymmetry of dependency between the gig platforms and the gig workers—one controls the ICTs and AI algorithms, while the other is easily replaced as workers in its execution and more open to exploitation (Heeks, 2022; Hickey & du Toit, 2013). The imbalance of power between system owners and users is evident when users provide their cultural and traditional ecological data to the system, while there is no established mechanism for reciprocal transparency from system owners. As a result, the benefits of digital systems in the Global South tend to accrue disproportionately to high-income countries in the Global North, such as the United States, China, and nations in Europe (Taylor & Broeders, 2015). This is true in Global North as well, where the super-rich extort this wealth from all of us in one way or another (e.g. published works scraped from the internet).

As technology is developed and becomes more dominant in the Global North, digital platforms that find their way to the Global South learn that their content is often not inclusive of diverse languages and cultural contexts of marginalised groups, where the Global North's languages and cultural preferences dominate and access costs are high and lock people into expensive contracts that determine their availability, use, and content. This digital injustice enables an aggregation of power well beyond what is feasible in most non-digital systems (Heeks, 2022). The digital divide and its negative impact on the socio-economic, health, and education conditions of those on the margin are contributing to the exploitation of the Global South's labour, cultural, and social capital, highlighting the power differentials between the Global North and Global South. The arc of colonialism continues to dominate North–South relations.

Gender, Race, and Indigenous Inequality

Gender and racial disparities are also a cause of concern. Education, income, and occupation explain in part the racial and gender divide in the use and access to ICTs. This divide isn't just about who has internet access, but it's also about how technology is used, the quality of that access, the opportunities it enables or restricts, and the representation of voices it includes. It is about prejudices and biases embedded in algorithms. It typically overlooks questions of power and inequality and fails to recognise that technologies are socially shaped by culture. Many technological systems—such as audio-visual platforms, online language translation tools, and algorithmic recruitment software—routinely fail to adequately recognise the needs, rights, and accomplishments of women, racialised groups, and other diverse communities (Arora et al, 2023). These digital and textual representations often perpetuate exclusion, misrecognition, and systemic bias, undermining equitable access and reinforcing existing social inequalities (Gupta, Parra & Dennehy, 2021). There is a positive correlation between internet connectivity and economic growth, individual empowerment, transformation, and community development. The absence of fair and uniform access of ICTs results in unequal opportunities for individuals and communities, leaving women, racial minority populations, and Indigenous communities at a distinct disadvantage (Padole, 2024). Gender, racial, and Indigenous disparities in use and access are presented as the new face of inequality limiting these population's opportunities to meaningfully and safely access its use and benefits. AI surveillance and tech-based coercive control, as well as cyberstalking and image-based abuse severely undermine autonomy, safety, and mental health for women, especially the disproportional violence faced by Indigenous women, and predictive policing can exert punitive control over marginalised groups such as young Muslim boys, while Indigenous, and remote communities are particularly susceptible to invisibility and lack of representation. These harms are often cumulative, persistent, and difficult to escape. Reversing their invisibility requires deliberate action through structural change.

Deprivation Implications

Commentators caution against reducing the digital divide to purely a technical and access issue as it overlooks the socio-cultural, educational, health, and political consequences (Cong & Quinn, 2025; Sanders & Scanlon, 2021). As many social, health, civic, economic, and political activities become digitalised, the lack of access and technology can hamper socio-economic benefits for those excluded and make the divide even wider (Yadov, Yadov & Pawer, 2025). We can see from the discussion above that the digital divide prevents and excludes people from participating in civic life, its rewards and benefits and at times leads to discrimination, social, and technological exclusion. It undermines economic growth and hinders inclusive development worldwide (Molala & Makhubele, 2021). It can reinforce and amplify existing harmful gender, racial, Indigenous stereotypes, and prejudices, trapping inherent biases and injustices into communications systems and

algorithm decision-making and further widening the digital divide and marginalising the poor and vulnerable even more (Yadav, Yadav & Pawar, 2025). Each deprivation reinforces the other.

Given how widespread ICT and AI technologies are, and their importance in accessing jobs, health care, and education, universal access to these tools having universal access it is argued offers the potential to empower individuals, foster economic growth, and promote social progress, ensuring that everyone can succeed in the digital era without facing bias or exclusion (Orellana-Rodriguez, 2024). It is now seen as a basic human right.

Digital Human Rights

The concept of digital human rights is not new; it builds on longstanding UN Human Rights Charter (1948a) evolving dramatically in response to technological change. The UN has launched many global initiatives to address the growing digital divide. These initiatives have been focused on infrastructure, affordability, digital literacy, and inclusive, reflective, and intersectional governance.

The UN High Commissioner for Human Rights has called the digital revolution a major global human rights issue, urging regulation and ethical oversight of AI and data systems. In 2024 the world leaders drafted Global Digital Compact with the aim to lay down shared principles and guidelines for how global technologies should be governed. Particularly stressed is the need to close all digital divides and accelerate progress across the Social Development Goals (SGDs). In essence, it calls on global community leaders to support the development required for an inclusive digital economy; to build an inclusive, open, safe, secure digital space for all; to strengthen international data governance and govern AI for all humanity (Global Digital Compact, 2024). UN-Habitat Digital Inclusion Playbook (2021) recognises that beyond providing access to infrastructure, support for digital skills and attitudes is crucial for inclusive and equal participation. Another important aspect included in the Playbook is choosing sustainable finance models and sustainable technological practices.

The implementation of legal rights to access ICTs and AI tools mustn't be just declarative. International constitutional protection and advocacy are one thing, but their implementation is another. It means that ICTs need more investment in large-scale infrastructure hardware such as cables, satellites, devices, internet providers, electricity as well as modem-router combos, and broadband access points. It means the state has obligations to promote access to the internet and digital technology for all its citizens in various contexts-social, economic, technological, personal, security, educational, and organisational and eliminate barriers (Bieliakov et al., 2023; Cong & Quin, 2025). These conditions can sometimes only be fully implemented over time, given their complexity and the need for substantial budgetary resources. In the meantime, policies can be enacted to protect citizens against geo-location abuse, support fair and healthy access, and ensure safe working conditions and work-life balance in the digital environment (Bieliakov et al., 2023). States can collaborate with major technology companies to enhance

network speed, reliability, and quality in remote and underserved regions, ultimately improving the affordability, coverage, and overall quality of global internet bandwidth. Other initiatives can be implemented immediately, such as promotion of digital literacy and formation of digital skills in different age groups, eliminating barriers to the internet caused by structural inequality, and poverty, uneven infrastructure, and internet connection quality. The list of responses also includes the creation of free access to the internet in public spaces and economic support for vulnerable people via subsidies or tariffs to ensure participation and access such as communal SKYLINK access via remote libraries (Bieliakov et al., 2023; 196). In Australia, as an example, many state libraries provide free internet access. Everyone has a right to access the internet and make free use of it if they have a library nearby.

Possessing and improving digital access and tools requires equipping individuals with skills to use them effectively and safely is equally critical. This requires individuals to understand online safety, undertake critical questioning of data and content, and have the skills to interpret, analyse, and utilise information effectively. This also requires users and potential users to become digitally literate. Digital literacy, as an example, requires people to have the critical thinking required to protect themselves from large-scale misinformation and disinformation. Digital literacy also includes the ability to locate, seek and use information to successfully shape the process of lifelong learning (UNESCO 2018), and identify and challenge lack of access and affordability and the existing digital divide.

Importantly encouragement for user-generated content to ensure digital content is accessible in multiple languages and formats and commitment to inclusivity is essential (Orellana-Rodriguez, 2024). Finally, collaboration between governments, NGOs, private entities, UN, and intergovernmental organisations can leverage resources, expertise, and pressure to create a collective responsibility to create a more equitable and inclusive digital future and develop innovative and affordable access solutions.

If ICTs and governments do not act on these suggestions, the digital divide will likely worsen unemployment, discrimination, and limit access to health care and education for marginalised groups. Ignoring this issue neglects a core social work responsibility: advancing social and human rights justice for those facing exclusion.

Social Work and the Digital Divide

As a human rights-based profession that promotes inclusion rather than exclusion, social work is often left to deal with the consequences of poverty, inequality, structural oppression, the lingering impact of colonisation and settler violence—just to mention a few of the current injustices social workers face day to day as daily practice is increasingly micro-managed, risk averse, and under-resourced. The latest social concern is with the unchecked proliferation of ICTs and their more pernicious impact by creating yet another impediment to equality and social, economic, and political participation by an ever-widening digital divide. As social workers are often on the front lines in addressing the needs of marginalised communities, they

are uniquely positioned to advocate for digital inclusion, equal access to resources and guide interventions aimed at addressing technological barriers (Sanders & Scanlon, 2021; Padole, 2024). Women and marginalised groups, including Indigenous communities, need to lead and design technology more suitable to their needs and wants for full civic, economic, and political participation.

Tackling this challenge is complex. AI systems are increasingly used in corporate, government, and social work settings as services move online; chatbots now assist with client interaction, counselling, education, and care. AI tools streamline tasks like admin, theory discovery, HR management, predictive analysis, and training. Addressing the digital divide remains crucial—social workers can use strategies to reduce technological gaps and prevent inequality (Yadav, Yadav & Pawar, 2025; Molala & Makhubele, 2021). Values such as respecting diversity and empowering individuals excluded by technology demand urgent focus. As ICTs and AI become essential post-pandemic, a political programme is needed to ensure technology advances support all, making professional activities more efficient and inclusive.

How to Respond? A Politics of Action

The politics of social work involve confronting injustice while seeking justice for all and have their roots in social activism and social change (Gray & Webb, 2013). Change through challenging the status quo involves both micro and macro efforts. As always, social work is about how the personal connects with the political. While social work's professional activities and resources are increasingly being digitalised provisions and actions must be made to ensure no one is left behind. To do this several levels of action need to be addressed simultaneously. This requires policies, programmes, and activism at Global, Regional, and Local levels aimed at closing the digital gap (Yadov, Yadov & Pawer, 2025; Padole, 2024; Molala & Makhubele, 2021; Sanders & Scanlon, 2021; Steyaert & Gould, 2009).

Global, Regional, and Local Advocacy

At the Global level, social work already has representation at the UN where its presence could encourage UN agencies as well as International Non-Governmental Organisations (INGOs) to continue to advocate for addressing the inequalities linked to the digital divide. This is possible as the International Association of Schools of Social Work (IASSW) has special consultative status with many influential UN agencies such as Economic and Social Council (ECOSOC), while the International Federation of Social Work (IFSW) works with UNHCR, UNICEF, WHO, and others to bring a social work perspective on global issues and promote humanitarian and developmental programmes worldwide. Initiatives could be to extend current efforts to include framing policies and programmes that support technology transfer to the Global South and promote affordable long-term digital financing policies for the developing nations in their pursuit of digital development (Yadov, Yadov & Pawer, 2025; Molala & Makhubele, 2021; Secretary general roadmap for digital cooperation 2020).

At the Regional level, social work can work through the Asian Pacific Association Social Work Educators (APASWE) with regional governments, NGOs, civic and business leaders to advocate for defensible and resilient ecosystems to address the digital divide at all levels of society. At the Local level, the Australia Association for Social Workers (AASW) and other groups like National Association of Social Workers (NASW) in the US can support the efforts of their constituencies to build capability through education, training, and capacity-building programmes and working to build digital infrastructure and policies that place people at the centre (Yadov, Yadov & Pawer, 2025).

Policy and Advocacy; National and Local

A policy and advocacy focus would include working with tech companies, NGOs, and intergovernmental entities to expand broadband and mobile network infrastructure in disadvantaged and underserviced regions and fight for digital inclusion in content and languages in all new technologies being developed. Reducing costs by pushing governments to provide subsidies and financial incentives for broader coverage and cheaper access for those impacted by the digital divide is another strategy (Sanders & Scanlon, 2021).

Promoting policies and advocacy to promote equal access to digital technologies can include:

1 Promoting free digital access in public spaces.
2 Collaborating with local and regional leaders to support digital inclusion efforts.
3 Promoting equitable distribution of digital infrastructure and ensuring that marginalised peoples and communities are considered in the rollout of technologies.
4 Mounting a public awareness campaign to pressure Big Tech to work with Governments, NGOs and civil society to create systems designed for inclusivity and equity, such as assistive technologies and digital content that are assessable in multiple languages and formats, and to design context-specific strategies for easier access.
5 Encouraging Big Tech companies to design local content; invest in women, Indigenous, and multi-ethnic-dominated value chains and sections to close the gender, racial, and Indigenous gaps; promote open-sourced tools and platforms that can be adapted locally; and ensure that applications and interface systems integrate culturally relevant examples and support local languages and dialects to enhance accessibility and user engagement (Bieliakovet al., 2023; Molala & Makhubele, 2021; Sanders & Scanlon, 2021).

Individual-Level Practitioner

In talking about the digital divide and preventing technology from becoming a new platform for social exclusion, individual practitioners must leverage their educational and community-building skills by:

1 Offering digital literacy workshops for different age groups and communities.
2 Facilitating peer mentorship programmes to encourage learning and confidence.

3 Pushing for access points in communities (libraries, community hubs with free internet and devices) access to telehealth services.
4 Advocating for home access with second-hand computers and open-source software to be available for those who do not have them.
5 Organising online social connections and virtual events to facilitate inclusion (Orellana-Rodriguez, 2024; Padole (Talekar), 2024; Molala & Makhubele, 2021).

In Conclusion

Access to technologies is a human right, essential for individual empowerment, economic growth, and social progress. Everyone should have the opportunity to thrive in the digital age (Orellana-Rodriguez, 2024; Sanders & Scanlon, 2021). Addressing the challenges of the multi-dimensional impact of the digital divide demands a collaborative approach involving governments; the private sector; civil society organisations such as INGOs, NGOs, grassroots movements, academic institutions, religious groups, and advocacy networks; intergovernmental organisations such as the UN and its many parts; and professional groups dealing with the social and welfare impact. We can't escape its impact, but we can enact strategies, policies, advocacy, and tactics that address its impact and attempt to close the digital divide as social work demonstrates a commitment to social justice and human rights for all.

References

Arora, M., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism and marginalisation. https://www.sciencedirect.com/science/article/abs/pii/S1471772723000325

Bailey, D. E., Faraj, S., Hinds, P. J., Leonardi, P. M., & von Krogh, G. (2022). We are all theorists of technology now: A relational perspective on emerging technology and organizing. *Organization Science, 33*(1), 1–18. https://doi.org/10.1287/orsc.2021.1510

beDots, O. (2025, January 26). The role of artificial intelligence in modern ICT. https://blog.bedots.io/2025/01/26/the-role-of-artificial-intelligence-in-modern-ict/

Bieliakov, K., Tykhomyrov, O., Radovetska, L., & Kostenko, O. (2023). Digital rights in the human rights system. *Intereulaweast, X, 1*, 183–207. https://doi.org/10.22598/iele.2023.10.1.10

Cong, Y., & Quinn, P. (2025). Bridging the digital divide: A human rights-based approach to digital literacy and equity in the EU. In S. Serpa & A. I. Santos (Eds.), *Digital equity and literacy* https://doi.org/10.5772/intechopen.1011614

Gray, M., & Webb, S. (2013). *The new politics of social work*. Palgrave Macmillan.

Gupta, M., Parra, C. M., & Dennehy, D. (2021). Questioning racial and gender bias in AI-based recommendations: DO espoused national cultural values matter? *Information Systems Frontiers*, *24*(1), 1465–1481. https://doi.org/10.1007/s10796-021-10156-2

Heeks, R. (2022). Digital inequality beyond the digital divide: Conceptualizing adverse digital incorporation in Global South. https://www.tandfonline.com/doi/full/10.1080/02681102.2022.2068492

HEPA. (n.d.). The role of ICT in the development of Artificial Intelligence. *HEPA Greece*. Retrieved from https://www.hepaoffice.gr/en/the-role-of-ict-in-the-development-of-artificial-intelligence/

Hickey, S., & du Toit, A. (2013). Adverse incorporation, social exclusion and chronic poverty. In A. Shepherd & J. Brunt (Eds.), *Chronic poverty* (pp. 134–159). Palgrave Macmillan.

International Telecommunication Union (ITU). (2023). *Measuring digital development: Facts and figures 2023*. Geneva: ITU. https://www.itu.int/itu-d/reports/statistics/facts-figures-2023

Molala, S. T., & Makhubele, J. C. (2021). The connection between digital divide and social exclusion: Implication for social work. *Humanities & Social Sciences Reviews*, *9*(4), 194–201. https://doi.org/10.18510/hssr.2021.9427

Orellana-Rodriguez, C. (2024). Understanding the digital divide in the age of AI: Where is the gap? https://www.libreai.com/digital-divide/understanding-the-digital-divide-a-comprehensive-exploration/

Padole (Talekar) A. (2024). Inequality in the digital age: The role of social work in bridging the digital divide. https://www.ijnrd.org/papers/IJNRD2412214.pdf

Report of Secretary-General: Roadmap for digital cooperation (2020). https://www.un.org/digital-emerging-technologies/content/roadmap-digital-cooperation

Sanders, C. K., & Scanlon, E. (2021). The digital divide is a human rights issue: Advancing social inclusion through social work advocacy. *Human Rights Social Worker, 6*(2), 130–143. https://doi.org/10.1007/s41134-020-00147-9

Steyaert, J., & Guld, N. (2009). Social work and the changing face of digital divide. *BJSW, 39*, 740–753. https://doi.org/10.1093/bjsw/bcp022

Taylor, L., & Broeders, D. (2015). In the name of development: Power, profit and the datafication of the Global South. *Geoforum, 64*, 229–237. https://doi.org/10.1016/j.geoforum.2015.07.002

UN – Habitat; Addressing the digital divide. (2021). https://unhabitat.org/addressing-the-digital-divide

UNESCO. (2018). A global framework of reference on digital literacy skills for indicator https://unesdoc.unesco.org/ark:/48223/pf0000265403

United Nations. (2024). *Global Digital Compact: A comprehensive global framework for digital cooperation and governance of Artificial Intelligence*. United Nations Office for Digital and Emerging Technologies. Retrieved from https://www.un.org/global-digital-compact/en

United Nations Human Rights charter. (1948a). https://www.un.org/en/about-us/universal-declaration-of-human-rights

van Dijk, J. (2020) *The digital divide*. Polity Press.

Yadav, R., Yadov, B., & Pawar, M. (2025). *The International Journal of Community and Social Development*, June, 222–239. https://doi.org/10.1177/25166026251350522

6 Critical Scholarship in the Age of Large Language Models

Russell Shuttleworth

Introduction[1]

The recent introduction of text-focused, generative artificial intelligence (GAI) has been met by academics with a mixture of excitement and concern about GAI's integration into the regimes of scholarship. In this chapter, I argue that scholars in the humanities or social sciences who deploy critical theory, broadly construed, should be cautious in their engagement with GAI because they represent an ethical threat to critical scholarship.[2] A critical theory of artificial intelligence (AI) must be concerned with structural biases underlying the development and design of these systems and their disempowerment of various marginalised groups (Lindgren, 2024; McQuillan, 2022). However, since large language models (LLMs) are increasingly being incorporated within academia, it is paramount that critical scholars also interrogate them in terms of possible ethical implications for their own scholarship.

To provide background for my argument, I discuss multiple issues involved in the rise and application of GAI/LLMs and deploy some ideas from critical theory, especially those of Cornelius Castoriadis. Castoriadis was centrally concerned with the alliance between technology and capitalism and techno-capitalism's unlimited expansion in modernity. He cogently elucidates the interpenetration of logic and meaning operating within this social imaginary, which he viewed as a challenge to critical autonomy.[3] Castoriadis's ideas provide food for thought on how critical scholars can contend with the introduction of LLMs into the academic mix, a technology that threatens their ethical relation to scholarship.

I further argue Castoriadis's elucidation of the project of autonomy can provide a useful aspirational narrative to reaffirm the creative space for critical theorising, assist critical scholars in navigating cultural and academic pressures to fully integrate LLMs into their scholarly practice, and reflexively interrogate their ethical engagement with this technology. I finally present several ethical relations to scholarly practice that may be useful in the age of LLMs.[4] An existential query for critical scholars operates in the shadows of this chapter, which I occasionally bring into relief: whether designers of GAI or artificial general intelligence (AGI) will be able to simulate a human intelligence capable of rendering a truly creative critique? No definitive answer is provided herein.

DOI: 10.4324/9781003627920-6
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

Definition, Research and Policies around Use of LLMs in Academic Scholarship

GAI employs machine deep learning which develops algorithms that by leveraging statistics identify and imitate correlations and patterns from large data sets (Banh & Strobel, 2023). LLMs, a GAI application, generates relevant text by predicting the most likely token to occur next given previous tokens. Different from discriminative AI based on variance, LLM's calculate probabilistic patterns from vast training datasets scraped from social media and/or third-party websites, and they respond to user prompts with a range of valid responses; they can quickly produce an essay if prompts dictate topic and criteria (Bahn & Strobel, 2023). Importantly, LLMs "have been programmed to recursively improve their performance" (Noonan, 2025, p. 132). GAI has been described as "a paradigm shift" (Pescape, 2024, p. 37) and is increasingly employed across a range of media (textual, visual, audio, etc.).

Research has found LLMs have several advantages for scholarship, including assisting with grammar, sentence structure, clarity, research tasks, efficiency, idea generation, early drafts, and alleviating workloads (AlZaabi et al., 2023; Liu et al., 2025; Rahman et al., 2023). Problems include generating inaccurate text or references, hallucinating and concerns about privacy, authorship, copyright, plagiarism, accountability, transparency and biases (e.g. race, gender) (AlZaabi et al., 2023; Liu et al., 2025; Rahman, et al., 2023). In terms of critical thinking, although there are some mixed and positive results for using LLMs in particular learning contexts (e.g. Patac & Patac Jr., 2025; Toma & Yanez-Perez, 2024), multiple studies indicate that while sometimes increasing efficiency, consistent use can atrophy critical thinking (e.g. Kosmyna et al., 2025; Tarisayi, 2024).

Academic publishers and university research ethics committees have responded to the various problems by developing ethical guidelines on responsible use. An informal review of GAI/LLM authorship/research policies revealed several points. Attributing authorship to LLMs is not permissible. Most policies, especially for research ethics, explain risks around confidentiality, bias, etc. Performing simpler research tasks like automating routine data collection is often permissible, more complex tasks like analysis not as much. Generative editing and/or autonomous content creation are/is often permissible; some policies state that LLM use should not replace human intellectual reasoning. Policies emphasise authors must ensure transparency, accuracy, validity and responsibility and provide oversight. Despite the ostensible misuse of LLMs, an underlying sense of optimism is apparent. Phrases such as 'scientific advancement', 'continuous improvement', and 'improve efficiency' are scattered throughout. These institutions, in short, are implying "the potential of LLMs to advance scientific progress: a primary good of academia" (Mann et al., 2024, p. 1272).

Even with the roll-out of policies, many individual scholars are nevertheless ethically concerned about their scholarly practice. In fact, the assumption that using LLMs for autonomous content creation is positive because it assists scientific progress, does not mesh well with critical scholarship in the humanities and social sciences. Indeed, critical scholars have reason to be sceptical when progress

is invoked to justify their use. For critical theorists, interrogating the complexity of social issues and critiquing power relations around their manifestation requires a different sensibility than a positivist sense of cumulative scientific knowledge acquisition and progress.

Critical Theoretical Implications of AI/GAI/LLMs

Prioritising Ensidic Logic and Rational (Non)Mastery in the Imaginary Signification of Progress

Similar to interpretivist theories, critical theory recognises the mutability and transformation of a society's meanings or linguistic signifieds, what Cornelius Castoriadis refers to as the open-ended "bundle of referrals" that is "constitutive of a living language" (Castoriadis, 1987, p. 345). Critical theory is further concerned with critiquing power relations around these changing meanings. Different from a neutral objectivity, critique nevertheless necessitates a certain distancing in which one interrogates the status quo and formulates alternatives. As Castoriadis (1997a, p. 36), emphasising the creative possibilities within signification and critique, puts it, "Critique entails a distance relative to the object…[and] presupposes the creation of new ideas, new standards, new forms of thought that establish this distance".

In contrast, both rationalists, who recognise the priority of human intellect, and positivists, who favour capturing experience through measurable, causal, predictive and/or probabilistic means, remain under the unreflexive spell of what Castoriadis (1987, 1997a, 1997b) terms ensemblist-identitarian logic (ensidic), the proto-organising logic underlying coding within discourse, language and technique. In this logic, sets and their elements are viewed as wholly determinable. Ensidic logic at its extreme is mathematics, but every spoken and written word is a manifestation of ensidic logic, as is the functional aspect in social doing (Castoriadis, 1987, 1997a, 1997b). An extreme emphasis on the application of ensidic logic is at the heart of the instrumentalisation and technologisation of human life in modern societies.

Critical theory has long critiqued the rise of instrumentalism. For example, the Frankfurt School of critical theory argued that the Enlightenment is "when instrumental rationality truly comes to dominate, as means-end calculation is the kind of reasoning required for capitalist production and efficient bureaucracy" (Celikates & Flynn, 2023, p. 16). Castoriadis adds the concept of the unlimited expansion of rational mastery to the mix—a notion implicit in Aristotle's rationalism and Descartes's and Bacon's sense of humanity's mastery and possession of nature (Papadimitropoulos, 2018; Castoriadis, 1991). The unlimited expansion of rational mastery—in other words 'progress'—is a complex signification that infuses both technological and economic development in modernity (Papadimitropoulos, 2018; Mouzakitis, 2017; Castoriadis, 1991).

Castoriadis (1987) refers to core significations in a society such as 'justice', 'God' or 'economy', as social imaginary significations. This is not because they are necessarily fictional, but because, as he observes with regard to economy, it is a "central social imaginary signification which do[es] not refer to something but on

the basis of which a host of things are socially represented, reflected, acted upon and made as economic" (1987, p. 362). A society's imaginary is comprised of its shared meanings and values, that is, a magma of social imaginary significations and secondary significations, which are comprised of both ensidic logic and an imaginary dimension, which are indissociable in practice (Castoriadis, 1987).

Castoriadis considered both progress and capitalism as core social imaginary significations in modernity. In terms of capitalism, Smith (2014, p. 161) notes that Castoriadis saw it as "the first social formation to legitimate itself on the grounds of instrumental reason". Capitalism is a process whereby the conditions and means of accumulation, production, commerce, finance and consumption are efficiently and ceaselessly transformed via rational mastery or progress (Castoriadis, 1987).[5] A key aspect of progress is efficiency, which, as Ellul maintains, is deployed as the primary criteria and acts to generally de-symbolise pre-techno-capitalist rituals and resymbolise them as "concretized technical reasons" (Nikoletos, 2024, p. 49).

Capitalism unsurprisingly relies heavily on technological advancement, which especially gathered steam during industrialisation when the artisan trades were supplanted by mass production of everyday items such as clothing during the 19th century (McQuillan, 2022). In this case, progress was achieved by effacing human artisan skills with a more efficient technology to expand capitalist markets. The capitalistic application of instrumentalism and technological development has progressively expanded toward mastery of individual and group dynamics and was given sharp impetus with the advent of what has been termed digital capitalism. With the proliferation of social media, Big Tech companies saw profitability in online activity and resorted to psychological and behavioural manipulation of consumers. Profitability also drives the hyper-promotion of GAI technology. LLMs promise to more efficiently produce text than its consumers. In this case, rational mastery aims to possess and simulate a skill, which over the past several centuries has come to at least partially denote human intelligence. Castoriadis (1991, pp. 272–273) commented thusly on the march of progress via rational mastery:

> As soon as 'rationality was seen to be perfectly 'objecifiable' (which quickly came to mean: capable of being put in the form of algorithm and susceptible to impersonal mastery), 'rational mastery' came to imply, and in truth requires impersonal mastery. But impersonal mastery, when extended to everything, obviously becomes the mastery of Nobody, of outis—and it becomes, thereby, complete non-mastery, impotence.

Castoriadis's rendering of impersonal mastery/non-mastery foreshadows the following statement from an editor of an online gaming magazine on the current state of GAI:

> anonymization is the point of generative AI. It's the logic behind the formation of datasets. Years of individual…thought are stacked and melted into a monolith, through which abstract traits and tendencies run like seams of vital

> ore. Individual labours are genericised to verbs cleansed of subjects, so… they can be appropriated and amalgamated into the predictive output.
>
> (Evans-Thirlwell, 2025)

Ullrich (2023), like Castoriadis, glimpses a logic of non-mastery in techno-capitalism, especially in the "computational paradigm" of GAI. This kind of non-mastery, he observes, "still occurs at the hands of humankind, with the only difference now being…ascribing responsibility for the destruction it may cause is shielded by a kind of autonomous limitlessness of its own creative possibilities" (p. 11). The masking of ethical accountability can indeed be witnessed in the current media frenzy over the supposed unbridled creativity of this new technology.

Simulating Creativity

From the outset, the design of LLMs is ensidic, as their complex algorithms prioritise identification of probabilities for a range of linguistic features such as morphology and word embeddings. Within multiple layers of a neural network is a vast complexity of variables interacting, learning, and adjusting from prediction errors. The precise decision-making process involved in responding to prompts remains unknown, and in the parlance of the industry, constitutes a black box (Banh & Strobel, 2023; Chaudhari, 2025; Kumar, 2024). However, LLM responses, simulated significations, are to varying degrees meaningful, depending on prompter expertise, if not always reliable (e.g. hallucinations, incorrect citations).

Research suggests LLMs are good at generating diverse ideas, especially initially; however, as Chaudhari (2025, p. 2) observes, "ChatGPT is…an enhancement application to creativity, …it doesn't create novel concepts but rather synthesizes various information from its training data to compile creative ideas". This point is reinforced by Cropley's (2025) analysis of LLM design as probability distribution and correlation between efficiency and novelty. As he contends, there is a "trade-off …between effectiveness and novelty in token selection", which constitutes a "built in barrier to LLM creativity" (p. 3). The ceiling on LLM creativity (0.25/1.00) corresponds to the mean of human creative performance (p. 6). LLM output is thus limited by training end-date, incomplete semantic grounding in topical contexts, limited diversity of voices and the very design of LLMs (Bank & Strobel, 2023; Chaudhari, 2025; Kumar, 2024l; Cropley, 2025); any resultant simulated signification cannot be considered part of an open-ended process. As Franceschelli and Musolesi (2024) observe, LLMs manifest combinatorial creativity, not transformational creativity.

In his later years (he died in 1997), Castoriadis acknowledged that the simulation of human cognition might be possible, but he remained sceptical that AI would ever precisely replicate human desires, imaginative and creative abilities. As he maintains,

> the question is not to know if we will one day manufacture a computer that… [can] simulate the passions, but if we will ever be able to formalize what we know as desires and as passions and, above all, as imagination.
>
> (2011, p. 66)

For Castoriadis, the social imaginary is a dynamic mode of being operating via history, whose tension is strung between the instituted (reproducing old structures) and the instituting (creating new forms). While past knowledge is the background for the creation of innovative ideas, it cannot determine them. History is the "emergence of radical novelty, and…we definitely must postulate a power of creation, a vis formandi, imminent to human collectivity [and]…individual human beings" (2011, p. 59). The open-endedness of a society's imaginary in the emergence of new linguistic referrals is a central aspect of this creative power (Castoriadis, 1987). Although Castoriadis was averse to hermeneutic perspectives, Adams (2011) suggests that using the term interpretive or contextual creation would better serve Castoriadis's notion of innovative creation informed by, but not determined, by the sociohistorical past.

For its part, the individual psyche exists as radical imagination, a continual imaginative flux of representations, affects and intentions/desires. Operating at both conscious and unconscious levels, the radical imagination eludes precise ensidic operations, its most basic form is the dream, which is experienced as indistinct and only ever interpretable ex post facto (Castoriadis, 1987, 1997a). It thus becomes clear that what Castoriadis means by formalisation above is that he is sceptical that the radical imagination and social imaginary will ever be broken down into an ensidic combination of determinable elements. Castoriadis (1987, 1997b) asserts that, whether happening as part of history's incessant process or albeit less often, expressed individually, creation of novel forms does occur.

In contemporary society, the concept of creativity primarily remains within a narrowly construed techno-capitalist/consumer orientation (De Cock et al., 2013, p. 150). In addition, within language specifically, the as-of-yet closure of machine learning to the open-endedness of imaginary significations and inability to access the imaginative flux of the psyche also work against transformational creativity. Nevertheless, researchers are working tirelessly to crack this closure to open-endedness and to access the imaginative process in their efforts to develop (AGI) and more closely approximate human creativity (e.g. Hughes et al., 2024). In the meantime, working with LLMs collaboratively via conversational-like prompting has been suggested as the best way to maximise LLM creativity (Carrigan, 2025; Franceschelli & Musolesi, 2024).

The Dark Side of Techno-Capitalism and GAI/LLM Use

The ethics of GAI/LLM development and integration into society is highly controversial. The large scale scraping of individual and organisational information gathered by large tech corporations via web crawlers from third-party websites or social media is clearly an exploitation of public trust. (Tiedrich, 2024). Ullrich (2023, p. 10) calls this scheme by techno-elites, "the greatest intellectual heist in history". While policy initiatives around consent and intellectual property in countries such as the EU have developed a broad regulatory approach, other countries such as the United States have opted for minimal regulations (Harris, 2025; Tiedrich, 2024). A vast number of workers are necessary to code data but are generally poorly

compensated. (Noonan, 2025). The mega-cost of fuelling AI/GAI coding and usage (Hensell, 2024), as well as its negative impact on the natural environment, is increasingly being recognised (Holzapfel, 2023). AI/GAI also incorporates bias against marginalised groups from its design and training data (McQuillan, 2022; Lindgren, 2024). Cognitive and creative work previously accomplished by individuals is increasingly being offloaded to GAI/LLMs (Bilderback, 2025), with a resultant loss of white collar and professional jobs (Piscione, 2025). Although it may be, as in prior technological shifts, that jobs lost open-up in new areas (Lindgren, 2024), it appears that GAI/LLMs yet pose a substantial threat to work relying on cognitive and artistic skills. In retrospect, effacement of artisan skills in the 19th century may have been a progenitor of the recent push toward GAI/LLMs minimising that need for human labour in the arts (e.g. visual arts, music, creative writing) and various cognitive work, which may eventually include academic work (e.g. research, scholarship, education).

Castoriadis (1997a) viewed the latter part of the 20th century as a retreat from autonomy and critical reason towards sociopolitical and cultural conformity. Driven by reliance on capitalism and rational mastery via technoscience, this retreat has taken a decidedly post-literate turn in the 21st century (Tucker, 2009). Indeed, many people now consume information—much of it banal—via the internet and social media. In the shadow of the global rise in authoritarianism, anti-critical thinking and anti-critical theory sensibilities are drastically rising (e.g. Burbules, 2022; Watson, 2023), as is widespread diminishment in reading proficiency and enrolment in humanities programmes (Tucker, 2009; Heller, 2023). The interaction of these factors reveals an overall erosion of critical sensibilities. This is the sociohistorical situation in which techno-capitalist innovations of GAI/LLMs made their entry.

For critical scholars, this dark side of techno-capitalism and GAI/LLM use is an important context for interrogation and critique. Issues around corporate control, coder exploitation, ongoing job losses, design biases, and disempowerment of marginalised groups in the coding process should be relentlessly critiqued, and a more egalitarian AI made a goal (McQuillan, 2022). Ongoing environmental harm should also be strongly critiqued, and strategies for societal sustainability should likewise be proposed (Sustainability, 2025). The dark side necessarily includes taking seriously the more existential points implied above regarding efforts to approximate human imagination itself and to foster total algorithmic reliance in an all-enveloping non-mastery. All this taken together constitutes the background to critical scholars' own engagement with LLMs in the ethics of their scholarly practice. This is the case whether the scholar chooses some level of complicity or not.

Towards Critical Autonomy in Scholarship

In his guide for using GAI/LLMs in scholarly work, Carrigan (2025) describes current and potential future harms of these systems including the existential possibility scholars might eventually be replaced with GAI/LLMs. He thus argues that scholars should "Establish…a mode of engagement with GenAI which expresses and is informed by our professional judgement…[which] could be a bulwark against

the threat on the horizon" (p. 149). Carrigan's book can be read as a treatise on how scholars should collaborate with GAI/LLMs rather than passively use these programmes to do the primary thinking. He argues that scholars should create a specialist understanding of GAI/LLMs by which to "fortify the position of academic labour within the system" (p. 149). Carrigan presents an ethically engaged but largely integrative appraisal of GAI/LLMs, how scholars can develop useful strategies to work with them and hopefully survive. Sometimes, however, he works to disrupt his own argument about using professional judgement. For example, he notes the difficulty in referencing all the ideas of others employed in one's writing in today's complex media environment and questions the notion of the individual author (à la post-structuralism). Carrigan thus diminishes the ethical integrity of at least trying to accurately acknowledge or cite others' ideas.

With pressures to integrate GAI/LLMs into scholarly practice, Carrigan offers a way in which this can be approached ethically. However, a glaring omission is that he does not adequately differentiate between critical and other kinds of scholarship and thus does not fully grasp the stakes involved for critical scholars. The perspective of the present chapter is that critical scholarship in the humanities and social sciences should more explicitly recognise its unique value to the scholarly enterprise in the age of LLMs. For critical scholars, contending with a host of ethical dilemmas and balancing complicity against benefits not only requires ethical responsibility, but also a commitment to human agency that works against usurpation of their critical skills and values their own ability to create transformative forms of critique. Being cognizant of the ethical complexities involved and recognising the value of their critical scholarship might shore up critical scholars' confidence to confront current and future threats to their professional standing and their sense of making a meaningful contribution. What might be helpful for critical scholars in the uncertain environment is an aspirational narrative that reminds them of their calling and can assist in navigating the ethical and scholarly difficulties in the age of LLMs. Castoriadis's (1987, 1997a, 1997b) elucidation of autonomy as a critical project, a primary component of his sociopolitical theory, is an exemplary candidate to fit this bill.

For Castoriadis (1987, 1997a), autonomy is an imaginary signification, which is increasingly under threat. It is not about freedom to choose arbitrarily or make political decisions without considering others, but autonomy requires both subjective and political modalities—"a political project for an autonomous society made by autonomous individuals" (Castoriadis, 2011, p. 59). An autonomous person critically reflects on their psychic and socially derived impulses and attitudes (1987, 1997a). The political mode of autonomy requires a society expunge heteronomous thinking and critically interrogate the social institutions that sustain it. Both modalities ceaselessly engage in critical interrogation of self and social institutions. Within this critically autonomous space, individuals and groups work against determinist tendencies in society and history and, when necessary, generate creative alternatives (Castoriadis, 1987, 1991). These processes are enabled by a deliberative democracy, which creates institutions that most facilitate individual autonomy and their "effective participation in all forms of explicit power…in society" (1991, p. 173).

Castoriadis also maintained that the eschewing of limits by the modern autonomisation of techno-capitalism has led to a host of world problems. As he puts it, "An autonomous society…must know how to be self-limiting…. individually and collectively. And capitalist society today…is running into the abyss from every point of view because it…doesn't know how to be self-limiting" (2011, p. 16). The most egregious problem being the ongoing environmental destruction occurring since industrialisation. With the environmental movement, Castoriadis recognised an exemplary project of political autonomy trying to set limits and transform social meanings and values (Castoriadis, 1997b). As Nikoletos (2024, p. 53) puts it, Castoriadis saw this movement as "capable of transforming not only the values and orientations of modern society but also the psychosocial constitution and existential attitude towards Being". A lesson to be learned from such resistance movements is that transformation often entails posing limitations to and resisting the destructive tendencies of dominant imaginary significations; in the present case, resisting the assumed progress signified by using LLMs.

According to Castoriadis, autonomy arose with the Greek polis and re-emerged tentatively around the 12th century in Western Europe reaching its peak 1750–1950 (Castoriadis, 1997a). As Castoriadis (2020, p. 27) states, the criticality of autonomy can be discerned in, for example, "Herodotus saying the Persians are infinitely better than the Greeks and the Egyptians are wiser". Castoriadis opposes autonomy to heteronomy, which relies on an external instituting power as the source of decision-making authority (e.g. God, the king/emperor, nature, or market forces), which thus constitutes a closure of a society's imaginary. In societies or historical periods where heteronomy is dominant, questions can be posed about the management of a particular power, but the power itself is considered beyond challenge (Castoriadis, 2020).

Castoriadis project of autonomy has been critiqued on several accounts; most notably for its purported Western origins, and for a binary opposition between autonomy and heteronomy (e.g. Arnason, 2001; Klooger, 2009). The debates about this are complex and I cannot do justice to them here. Suffice to say, as Klooger (2009, p. 292) emphasises, the closure that Castoriadis attributed to heteronomous societies is "untenable". Different modes of self-reflexivity occur in many societies. The degree to which they emphasise the radical critique that Castoriadis attributes to the concept of autonomy in Western society is however a question for empirical research (Klooger, 2009). Arnason (2023, p. 180) acknowledges that while autonomy is still "useful as a capacity of explicit and unlimited interrogation", within the contemporary situation of multiple modernities, autonomy should become more nuanced and downgraded from a revolutionary project. Indeed, nuancing the autonomy/heteronomy binary and empirically researching how these concepts may manifest in different societies are important points to incorporate into contemporary critical theory. Castoriadis (2020, p. 39) himself admits that he perhaps overstated his case, "I want to shake people up…[and] make it understood that man is not, by divine right, a democratic being, that democracy was a creation and a conquest of history…[and] is constantly in danger".

Arnason also maintains that the project of autonomy provides a minimalist ethic. There are no guarantees—unforeseen meanings and entanglement between a plurality of values are certain (Adams & Arnason, 2022). Yet, there is already a well-developed ethics for academic integrity and scholarship. While acknowledging autonomy's (and heteronomy's) ideality as Castoriadis originally envisioned, and that reality is fraught with complexity, I contend that the crux of autonomy as creative self/societal critique is yet worthwhile to pursue for critical theorists. Employing the project of autonomy as an aspirational narrative provides a skeletal path to which an ethics for critical scholarship can flexibly aspire. The idea of critical autonomy, its emphasis on and valuing of continual critique of both subject and society, could fuel critical-reflexive evaluation of scholars' ethical engagement in their scholarly practice and of the institutions and social imaginary significations implicated in the current crisis of scholarship. However, this narrative and evaluation should not solely be concerned about the ethics of scholars' own use of LLMs—an ethical relation to critical scholarship is a pursuit shared by other critical scholars. So, it is imperative this project open-up to the community of critical theorists, to collectively discuss the stakes involved, to provide sober evaluations of any social and environmental harms due to the take-up of this new technology, and to weigh its usefulness against possible costs for the critical scholarly enterprise.

Critical Scholarship in the Age of LLMs

Media is constantly reporting that AI will revolutionise scientific discovery. While this news is often about natural science and medical research, there is often the implication that AI can solve sociopolitical problems such as inequality, social marginalisation, etc. Yet the problems critical scholars interrogate cannot be solved computationally because "they are not functions of computation…but the intentional strivings that define human social life" (Noonan, 2025, p. 149). Critical scholarship is not simply about acquiring knowledge for progress; it is about acquiring a critical relation to knowledge for critiquing power dynamics in societies. While critical theory often aspires toward an egalitarian transformation, critical scholarship cannot be considered to be working toward a *scientific* breakthrough. I have suggested that Castoriadis's project of autonomy can provide an aspirational narrative for critical scholars in the age of LLMs. This does not necessarily mean that LLMs should never be employed in scholarship, but any particular use must acknowledge current and potential social harms and evaluate how the integrity of critically autonomous scholarship was maintained. The following are several ways critical scholars might ethically relate to their scholarship but are not mutually exclusive.

One ethical relation to scholarship would be through an ethics of resistance—taking a stand against the social imaginary signification of progress with its criteria of efficiency. This would take direct aim at the production-line orientation of publishing in academia. The slow-scholarship movement occurred a decade before the advent of GAI/LLMs (e.g. Hartman & Darab, 2012), but its central message

becomes even more important as the efficiency meter is cranked up. After all, does the notion of efficient critique make any sense for critical scholars? Resistance to efficiency would obviously be at odds with current university management priorities, the obsession with metrics and accountability, and the recent push by universities for scholars and educators to embrace LLMs. However, in terms of research and scholarship specifically, one must ask, what does the LLMs selling point of saving time mean? It obviously synchronises well with producing additional publications for the university's benefit. But this metric has never meshed with critical scholars' emphasis on quality over quantity. Carrigan (2025, p. 47) also argues against the instrumental use of GAI/LLMs; using them solely for efficiency "will make ethical practice difficult and even lead to the unravelling of academic authority".

Recognising the value of one's scholarly experience and skills would be another ethical relation to scholarship. A critical scholar's lifetime of reading, writing and theorising may not have as broad a scope as an LLM, but it is systematic to some degree, always idiosyncratic, and often serendipitous, which can result in multiple analytic possibilities and rich conceptual depth. A scholar might have an area of interest and several ideas to begin a project, but the creative dynamism of the psyche as radical imagination drives a continuous, moment-to-moment representative, affective, and desiring flux of images (Castoriadis, 1987, 1997b). The human imagination is fuelled especially by desire, a feeling of passionate interest to critically interrogate one or more topics, which is as-of-yet unsimulable by AI. In fact, the steps involved are not always explicitly planned, as the scholar reflects on, within and between multiple significations and their referrals or implications, finding new sources, going back to original sources, consulting colleagues, driven by a background of thoughts, feelings and, desire bent on critical interrogation, and creative critique. Even wrong-turns and mistakes can become creative in this process (Noonan, 2005). Writing multiple drafts is often about conceptualising, elucidating and sometimes expanding or opening up the meanings of particular social imaginary significations as they interact with a host of secondary significations, that is, thinking the argument/critique through to some kind of clarity. A long-time scholar's lived, embodied experience of reading, writing, and applying critical theory manifests as a formidable interrogatory-critical-analytic-interpretive-creative skill. Over-relying on LLMs, can short circuit this deep learning and creative process—banking on the broader scope of LLMs rather than the idiosyncrasy of a scholar's long intellectual journey. While LLMs can identify explicit patterns, themes or narratives in data sets, long-time critical theorists may be sensitive to latent meanings and counter-narratives, and their work can sometimes reach levels of scholarly depth that engender substantial innovation.

Providing a critical scholar has a solid foundation in researching society, reading extensively, critically interrogating, and writing, employing an LLM to generate text that selectively assists in scholarly tasks might be useful and constitutes another ethical relation to scholarship. An LLM could be inserted into the above process, especially if one has in-depth scholarly experience and can leverage it

to produce less broad and more nuanced prompts. However, this tact is rife with dangers; one being that early career scholars may be less able to formulate nuanced prompts. Key ethical concerns are how to maintain decision-making agency and a historical connection to and documentation of previous contextual ideas and theoretical concepts in the literature, as well as how to ethically claim authorship of the interrogative-critical-analytic-interpretive-creative rendering of the final critique. Selecting tasks less central to intellectual labour such as identifying sources of information, surveying texts or data for trends and patterns, and summarisation are relatively, but not always, lower risk. An additional step would be working with an LLM, in Carrigan's terms viewing it as an "interlocutor", through progressively refined prompts to develop substantive insights.[6] It is easy, however, for accountability to lapse in this process. Insightful responses should be queried for their sources, which are not always forthcoming. Scholars should ground their analyses-interpretations-critiques in a context of original sources when possible. Concepts and theories suggested should be followed up and comprehended, so scholars can ethically claim understanding. Evaluation of a critical scholar's ethical relation to their scholarship would require that they have a detailed understanding of the contextual ideas and major concepts they decide to include in their theoretical mix and be able to clearly explain conceptual linkages between the various ideas they deploy in their critique. Transparency around this process is a minimal requirement regardless of whether scholars collaborate with LLMs or not.

Conclusion

I have argued that a committed ethical relation to critical scholarship would benefit from employing Castoriadis's project of autonomy as an aspirational narrative through continual reflexive interrogation of their scholarly practice in the age of LLMs. This is not about providing an inflexible pathway, but to assist critical scholars in maintaining a sense of ethical awareness and also of agency in scholarly work and to value their critical interrogatory skills and ability to create forms of critique that are innovative and sometimes transformative. Against media and academic contexts that in their different ways promote increasing use of GAI/LLMs, critical scholars need to weigh any scholarly benefits of employing LLMs against opposing factors such as social and environmental impacts and detriments to ethical and creative scholarship. It is also imperative critical scholars explicitly recognise why they might have a different sensibility regarding using LLMs than scholars who work towards universal truth and scientific progress. Not to oppose science where it is useful, but to recognise the stakes involved for critical scholarship. It is also important to nurture a community of critical scholars committed to maintaining an ongoing discussion on the ethics of LLMs. While using LLMs can be justified if ethically defensible, wholesale employment by critical scholars would transgress scholarly limits and would sacrifice their critical autonomy on the techno-capitalist pyres of efficiency and progress and at best result in a productivity lacking substantial humanity.

Notes

1 I want to thank the two reviewers for their critical feedback on this chapter, as well as John Fox and Bob Pease for their useful comments on an early version.
2 By critical scholarship I mean scholarship that derives its conceptual framework from critical theory broadly construed.
3 I employ quite a few of Castoriadis's concepts in this chapter. For space and accessibility, many relevant concepts have not been included.
4 Due to space limitations, pedagogical issues will not be covered in this chapter. See Straume (2017) and Ablett and Morley (2020) for pre-COVID applications of Castoriadis's theory to pedagogy.
5 Smith (2014b, p. 176) has made a strong case for a more flexible perspective: capitalism produces "distinctive variet[ies] of accumulation", which push back against rational mastery and "coexist in permanent tension" with autonomy, modifying both.
6 See Carrigan (2025) for specific strategies.

References

Ablett, P., & Morley, C. (2020). Critical social work education as democratic paideia. In C. Morley, P. Ablett, C. Noble, & S. Cowden (Eds.), *The Routledge handbook of critical pedagogies for social work* (pp. 176–188). Routledge.

Adams, S. (2011). *Castoriadis's ontology: Being and creation*. Fordham.

Adams, S., & Arnason, J. (2022). A further conversation on social imaginaries: Political philosophy, normative commitments and the creativity of social action. *International Journal Social Imaginaries*, *1*, 328–353.

AlZaabi, A., ALamri, A., Albalushi, H., Aljabri, R., & AalAbdulsallam, A. (2023). ChatGPT in academic research: A review of benefits, concerns, and recommendations. *bioRxiv,* 2023-08.

Arnason, J. (2001). Autonomy and axiality: Comparative perspectives on the Greek breakthrough. In J. Arnason & P. Murphy (Eds.), *Agon, logos, polis: The Greek achievement and its aftermath* (pp. 155–206). Franz Steiner Verlag.

Arnason, J. (2023). Lessons from Castoriadis: Downsizing critical theory and defusing the concept of society. *European Journal of Social Theory, 26*(2), 180–200. https://doi.org/10.1177/13684310221117353.

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets, 33*(63), 1–17 https://doi.org/10.1007/s12525-023-00680-1

Bilderback, S. (2025). Can't work without it: The quiet addiction to AI at work. *Strategic HR Review*, 24(6), 227–232. https://doi.org/10.1108/SHR-05-2025-0045

Burbules, N. (2022). Promoting critical thinking in anti-critical thinking times: Lessons from COVID discourse. *Philosophical Inquiry in Education*, *29*(1), 5–10.

Carrigan, M. (2025). *Artificial intelligence for academics*. Sage.

Castoriadis, C. (1987). *The imaginary institution of society*. K. Blamey (Translator). MIT.

Castoriadis, C. (1991). Philosophy, politics, autonomy: Essays in political philosophy. D. Curtis (Ed.). Oxford.

Castoriadis, C. (1997a). *World in fragments*. D. Curtis (Translator & Ed.). Stanford University Press.

Castoriadis, C. (1997b). *The Castoriadis reader*. D. Curtis (Translator & Ed.). Blackwell.

Castoriadis, C. (2011). *Postscript on insignificance: Dialogues with Cornelius Castoriadis*. G. Rockhill & J. Garner (Translator). Continuum.

Castoriadis, C. (2020). *Democracy and relativism*. J. Garner (Translator). Rowman & Littlefield.

Celikates, R., & Flynn, J. (2023). Critical theory (Frankfurt School). In *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/critical-theory/

Chaudhari, J. (2025). Authorship and originality in the age of AI: A literature review on ChatGPT creative writing applications. *The National High School Journal Science*, 1–6.

Cropley, D. (2025). "The cat sat on the…?" Why Generative AI has limited creativity. *The Journal of Creative Behavior, 59*(4). https://doi.org/10.1002/jocb.70077

De Cock, C., Rehn, A., & Berry, D. (2013). For a critical creativity: The radical imagination of Cornelius Castoriadis. In K. Thomas & J. Chan (Eds.), *Handbook of research on creativity* (pp. 150–161). Elgar.

Evans-Thirlwell, E. (2025). We need to reclaim software's wishing well from the cruelty of generative AI. *Rock Paper Shotgun*, July 30. https://www.rockpapershotgun.com/we-need-to-reclaim-softwares-wishing-well-from-the-cruelty-of-generative-ai

Franceschelli, G., & Musolesi, M. (2024). On the creativity of large language models. *AI & Society*, *40*, 3785–3795.

Harris, L. (2025). Summary: Regulating artifical intelligence: U.S. and international approaches and considerations in Congress. *Congressional Research*, R48555, June 4. https://www.congress.gov/crs_external_products/R/PDF/R48555/R48555.2.pdf

Hartman, Y., & Darab, S. (2012). A call for slow scholarship: A case study on the intensification of academic life and its implications for pedagogy. *Review of Education, Pedagogy, and Cultural Studies*, *34*(1–2), 49–60.

Heller, N. (2023). The end of the English major. *The New Yorker*, February 27. https://www.newyorker.com/magazine/2023/03/06/the-end-of-the-english-major

Hensell, W. (2024). The billion-dollar price tag of building AI. *Time*, June 3. https://time.com/6984292/cost-artificial-intelligence-compute-epoch-report/

Holzapfel, A. (2023). Introducing political ecology of Creative-AI. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 691–704). Edward Elgar.

Hughes, E., Dennis, M., & Parker-Holder, J. (2024). Position: Open-endedness is essential for artificial superhuman intelligence. In *Proceedings of 41st International Conference on Machine Learning*, Vienna. PMLR 235.

Klooger, J. (2009). *Castoriadis, psyche, society, autonomy*. Brill.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X., Beresnitizky, A., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI Assistant for essay writing task*. MIT Media Lab.

Kumar, P. (2024). Large language models: Survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, *57*(260). https://doi.org/10.1007/s10462-024-10888-y

Lindgren, S. (2024). *Critical theory of AI*. Polity.

Liu, Y., Kong, W., Merve, K. Chat GPT (2025). Applications in academic writing: A review of potential, limitations, and ethical challenges. *Arquivos Brasileiros de Oftalmologia*, *88*(3), S0004-27492025000300700. https://doi.org/10.5935/0004-2749.2024-0269

Mann, S., Vazirani, A. A., & Aboy, M. (2024). Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nature Machine Intelligence, 6*, 1272–1274. https://doi.org/10.1038/s42256-024-00922-7

McQuillan, D. (2022). *Resisting AI: An anti-fascist approach to artificial intelligence*. Bristol University.

Mouzakitis, A. (2017). Modernity and the idea of progress. *Frontiers Sociology, 2*(3). https://www.frontiersin.org/journals/sociology/articles/10.3389/fsoc.2017.00003

Nikoletos, N. (2024). Cornelius Castoriadis and Jacques Ellul on the dilemmas of technical autonomy. *Thesis Eleven*, *182*(1), 41–56.

Noonan, J. (2025). Artificial intelligence, alienation and the existential conditions of human flourishing. In A. Bielskis (Ed.), *Human flourishing in the age of digital capitalism: AI, automation and alienation* (pp. 129–158). Bloomsbury.

Papadimitropoulos, V. (2018). The rational mastery in the work of Cornelius Castoriadis. *Capitalism Nature Socialism, 29*(3), 51–67.

Patac, L., & Patac, A. (2025). Using ChatGPT for academic support: Managing cognitive load and enhancing learning efficiency – A phenomenological approach. *Social Science & Humanities, 1* https://www.sciencedirect.com/science/article/pii/S2590291125000282

Pescape, A. (2024). Exploring the current state and future potential of generative artificial intelligence. In F. Santoianni, G. Giannini & A. Ciasullo (Eds.), *Mind, body, and digital brains* (pp. 37–56). Springer.

Piscione, D. (2025). The shrinking organization due to AI: The coming white-collar blood-bath (May 31). White Paper. Available at SSRN: https://ssrn.com/abstract=5277389 or https://doi.org/10.2139/ssrn.5277389

Rahman, M., Terano, H., Rahman, N., Salamzadeh, A., & Rahman, S. (2023). ChatGPT and academic research: A review and recommendations based on practical examples. *Journal of Education, Management and Develpoment Studies, 3*(1), 1–12. https://doi.org/10.52631/jemds.v3i1.175

Smith, J. (2014a). Capitalism. In S. Adams (Ed.), *Cornelius castoriadis: Key concepts* (pp. 155–166). Bloomsbury.

Smith, J. (2014b). Contexts of capitalism: From the "unlimited extension of 'rational mastery' to civilizational varieties of accumulation and economic imagination. In V. Karalis (Ed.), *Cornelius castoriadis and radical democracy* (pp. 158–176). Brill.

Smulewicz-Zucker, G. (2017). The Frankfurt School and the critique of instrumental reason. In M. Thompson (Ed.), *The Palgrave handbook of critical theory. Political philosophy and public purpose* (pp. 185–206). Palgrave.

Straume, I. (2016). Democracy, education and the need for politics. *Studies Philosophy Education, 35*, 29–45.

Sustainability Directory (2025). Ethical critique. Sustainability Directory. https://lifestyle.sustainability-directory.com/term/ethical-critique/

Tarisayi, K. S. (2024). ChatGPT use in universities in South Africa through a socio-technical lens. *Cogent Education, 11*(1), 1–10. https://doi.org/10.1080/2331186X.2023.2295654

Tiedrich, L. (2024). Global partnership AI (Organisation for Economic Co-operation and Development) Blog https://oecd.ai/en/wonk/data-scraping-responsibly

Toma, R., & Yanez-Perez, I. (2024). Effects of ChatGPT use on undergraduate students' creativity: A threat to creative thinking? *Discover Artificial Intelligence*, *4*(74). https://link.springer.com/article/10.1007/s44163-024-00172-x

Tucker, P. (2009). The dawn of the Postliterate Age. *The Futurist*, November-December. https://www.wfs.org/node/1000

Ullrich, C. (2023). The singularity of AI? Toward a new philosophical-theological intelligence. *Stellenbosch Theological*, *9*(1), 1–17.

Watson, L. (2023). "The Anti-Critical Race Theory" campaign—Classroom censorship and racial backlash by another name. *Harvard Civil Rights-Civil Liberties Law Review, 58*, 487–549.

7 AI and the Corporate University

Constructing New Imaginaries or Reinscribing Old Stratifications?

Sharon Aris and Georgia Van Toorn

Introduction

The release of ChatGPT in November 2022 propelled artificial intelligence technologies into the spotlight of higher education debates around technological innovation, pedagogy, and academic integrity. In Australia, as elsewhere, universities have leaned into Generative AI's (GenAI) promissory narratives and growing techno-hype, investing significant financial and human resources into data-driven teaching, learning, and communications platforms, administrative automation, research support, and student services. Yet rarely have these moves been subject to the kind of robust debate we typically associate with university environments. As these institutions scramble to embrace GenAI amid efforts to differentiate themselves in a globally competitive marketplace for higher education, questions arise as to how their adoption of GenAI tools aligns with the core academic values and public purpose of the university.

This chapter explores how GenAI is disrupting higher education in ways that significantly (re)define the core values and public purpose of universities, and the future professionals they produce. By GenAI, we refer to the wide scope of AI systems capable of generating new content, including popular large language models (LLMs) such as ChatGPT, AI planning and decision-making tools like agentic AI and other automated decision-making (ADM) and machine-learning (ML) technologies that process data to generate outputs, make recommendations or inform actions (Crawford, 2021; Zajko, 2022). In this chapter, our focus is less on the technical, classroom-based applications of AI – covered extensively elsewhere (c.f. Bearman & Ajjawi, 2023) – and more on how universities publicly portray their adoption of AI, situating these representations within wider neoliberal governance logics operating in higher education (Connell, 2019). In particular, we examine how AI adoption sits awkwardly alongside other elements of universities' core mission including their role in cultivating among students a critical sensibility and the capacity to engage with and challenge dominant social and technological forces.

Our approach in this chapter is principally conceptual, informed by critical sociological approaches to AI (Benjamin, 2019; Eubanks, 2018). To provide an anchor for our discussion, we conducted a scan of the public-facing content relating to AI on Australian university websites. The content and discourses we

DOI: 10.4324/9781003627920-7
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

examined reflect how universities' priorities around AI are framed, and allow us to observe differences in these portrayals, including apparent stratifications in AI use and application that mirror patterns of institutional wealth and status. The discourses and disparities we identify raise questions about what we want public universities to be and do. Finally, we suggest how universities could lead more critical engagement with AI for more equitable, democratic and sustainable futures (Vicsek, 2021).

Universities in the Neoliberal Context

GenAI has emerged after a 40-year period of extended and significant change to universities. Over this time, Australian universities, in common with universities across much of the global north, have faced the intersecting pressures of declining government funding for research, greatly expanded local, and international enrolments and increased government regulation (Croucher, 2023; OECD, 2023). In response to their funding shortfall, universities have enrolled increasing numbers of international students who pay full fees (Department of Education, 2018, 2024a, 2024b) and who now form a primary source of income for many universities. At the same time, local enrolments have also expanded (albeit unevenly across disciplines), prompted by government policy targets for increased tertiary qualifications (Department of Education, 2024b). This massification of higher education has been accompanied by a transformation of conditions for university staff, who face historically high rates of casualisation, precarity, underpayment, and the erosion of secure academic pathways (Burton & Bowman, 2022). The combined characteristics of competitiveness, regulation, expansion, and individualisation can be recognised as typical of the corporatist or "neoliberal university" (Ball, 2012). In this model, universities are positioned primarily as service providers, with students their consumers, who expect to reap a 'return on investment' for their increasingly expensive degrees (c.f. Bok, 2009; Collini, 2012; Connell, 2019; Marginson, 2013). The effect of this can be seen in the Humanities, Arts, and Social Sciences (HASS) disciplines, which have experienced a reduction in enrolments over the past decade in Australia, in part due to changed study fee subsidies based on a perception (by policy makers and the general public) that they do not prepare 'work-ready' graduates, despite an employment rate superior to other STEM degrees (Academy of Social Sciences in Australia, 2021).

This has led critics like Collini (2012, 2017) to argue that universities have been misdirected from their core purpose of pursuing understanding through critical engagement with intellectual traditions. Rather, universities are being reconfigured in the image of Silicon Valley entrepreneurship, where technological innovation, efficiency, and profit reign supreme (Williamson, 2017). In this shift, university education is less oriented toward the cultivation of "the life of the mind" (Collini, 2017, p. 43) or the formation of well-rounded citizens, but rather toward producing future workers, whose learning is centred on acquiring the skills and competencies demanded by fast-paced, tech-centric economies (Zouhaier, 2023).

Universities in the Age of AI

The burgeoning literature on how universities frame artificial intelligence can be grouped around three focal points: the socio-technical drivers of digitalisation and datafication of universities, university responses to AI in relation to assessment integrity and pedagogic practice, and critical discourses on AI in higher education.

Even before AI, universities were highly digitised, with core functions from administration, teaching, and assessment to library access, and learning support services increasingly mediated through digital platforms (Komljenovic et al., 2025). A more recent wave of technological transformation has seen data analytics applied to student recruitment, retention, program management, and academic integrity in the name of greater efficiency and personalisation of service provision (Henderson et al., 2017). Universities are now increasingly dependent on private sector tech companies for what and how students learn, what and how academic research is conducted and who has authority to make decisions (Komljenovic et al., 2025, pp.282–283). These shifts in and around core university infrastructure mean that higher education itself has become increasingly embedded in the platform environments of 'Big Tech', which now extend to the domain of GenAI. For instance, Microsoft, IBM, Google, and other Big Tech platforms now partner frequently with many top universities, including but not only those that specialise in AI research and development (Shao et al., 2020). This embeddedness has been criticised for, among other things, promoting a 'blind faith' in data, leaving unproblematised the degree to which data itself is problematic, particularly as data biases often represent the interests and dominant status of privileged social groups, leading to discriminatory automated decision-making (Noble, 2018; Eubanks, 2018). Critics also point to the limited agency that individuals – particularly students – have when engaging with data (Selwyn, 2015) and the fact that 'analytics-driven education runs the risk of being reductive, non-representative, unjust, and exploitative' (Selwyn, 2019, p.15). Digital divides may also be deepening socioeconomic divisions, as universities with lower incomes do not have the same resources to invest in costly digital and AI development, meaning that some students may be falling behind in standards of digital literacy compared to their peers at more prosperous institutions (Hannan & Liu, 2023).

The academic literature on pedagogy and AI commonly highlights both its affordances for students while grappling with its perceived challenges and risks to academic integrity (Lee & Low, 2024; Lodge, 2024; Luo, 2024). University governance and regulatory responses to AI include proposals for top-down regulation with multi-stakeholder, collaborative oversight (Dotan et al., 2024) and tend to stress academic freedom and warn against overly technocratic, market-driven controls (Driessens & Pischetola, 2024). There is also considerable work on strategies aimed at forestalling AI generated content in assessment, with proposals such as risk-rating assessments, policies on declaration of AI use, as well as advocating for more comprehensive rethinking of assessment (c.f. Corbin et al., 2025). A parallel body of studies focuses on attempts to enhance student agency and critical engagement with AI (Lee & Low, 2024; Luo, 2024) and complementary to this is

a teaching-focused literature on how AI is reshaping pedagogical approaches in higher education. For instance, Henderson et al. (2017) argue that effective digital education must be grounded in pedagogy rather than technology, with educators exploring the use of AI for personalised learning, feedback automation, and curriculum design.

Studies examining university discourses on AI most frequently focus on the apparent inevitability of AI. These include AI being presented as 'the future', as a race in which universities must participate, or what Bearman et al. (2023) characterise as a discourse of 'imperative change', where AI is represented as an inevitable force which must be met with either a dystopian 'resist to survive' or utopian 'positive transformation' perspective. Similar positive or negative framings are also found in other papers, including earlier work by Selwyn (2014) which distinguished between tech-pessimist 'doomer' responses and tech-optimist 'boosters'. A scoping review examining university responses to AI found an overall positivity, with the authors framing it as a mechanism to boost existing pedagogic reform agendas including changes to assessment practices and a focus on student agency in learning (Jensen et al., 2024). Less optimistically, Bearman et al. (2023, p.369) describe a 'discourse of altering authority', which focuses on the change in agency and authority of academic work, whereby teachers are de-centred and authority is more widely spread between 'staff, machines, corporations or students'. That is, the relationships in universities themselves will change, but academics will likely have less central control.

These literatures suggest that AI's impact on the university emerges at the intersection of three dynamics: the increasing enmeshment of universities with the private educational technology (EdTech) sector, contests between AI regulation and AI affordance in universities that were already producing existential and social effects on universities, and a dominant discourse on the inevitability of digital change. Following this, we focus on examining the ways universities market their adoption of AI on public-facing websites, to understand how these narratives frame AI's role in higher education, and what promises or risks are emphasised.

Public Framings of AI in Education

With the aim to map AI uptake in Australian universities and how this uptake is publicly portrayed, we conducted a scan of 20 Australian university websites focusing on AI-related content. While website content cannot reveal the internal deliberations, intentions behind, or practical realities shaping AI adoption, examining this does nonetheless provide a useful lens for exploring, at an institutional level, universities' public-facing priorities and how they frame their engagement with AI.

Our scan (February to May 2025) purposively sampled websites to include equal proportions of the top-ranked, high-income, high-profile institutions featured at the top of university finance and performance lists, and smaller, regional universities more likely to feature down the bottom of such lists (drawn from Australian Government, 2022; Australian Financial Review, 2024; Edurank, 2025; UniversityRankings.com.au, 2024; Times Higher Education, 2025). Searching on university

websites using the keywords 'AI' and 'artificial intelligence' we charted in a table content from the first two pages of search results. From this, we developed key categories relating to how universities portrayed their use of AI including:

- In marketing/promotional materials.
- As teaching, research, and administration supports for staff.
- As learning aids for students.
- As subject to policies, ethical principles, resources, and rules governing academic integrity and student conduct.

We also charted more general content relating to AI, such as university position statements, announcements of AI-related initiatives, and strategic plans, to capture broader discursive framings and public-facing institutional priorities. Using this information, we first examined what characteristics all universities had in common, then identified key differences. These patterns guide the description and discussion below. While useful descriptions are made, our analysis is inevitably limited, as continuously updated web content produces constantly changing texts, meaning our findings represent only a selective snapshot in time (Kim & Kuljis, 2010).

The Socio-Technical University – Industry Development and Regulation of Student Use

The scan of AI content on Australian university websites revealed that while there is a wide variety of university innovations and responses to AI featured, two areas dominated: university emphasis on commercial AI development and application; and university regulation, support, and restriction of AI use by students. These two areas were consistent for both highly resourced urban universities and smaller regional ones.

All universities featured materials on their websites relating to AI technical innovation, research, and development. This included features on commercial AI development and research, AI-related educational offerings, and academic expertise in AI at a university, school, or individual level. One of the most common examples was the foregrounding of partnerships with commercial AI technology firms. The tone and focus of these pages were overwhelmingly positive, highlighting the affordances of AI in relation to forecasting, efficiency, and the scope and speed of information processing. A much smaller number engaged with critique or discussion of AI-related risks or social challenges.

Many, but not all, institutions also offered AI-specific courses or degree programs. These included post-graduate and undergraduate study programs, majors, AI-related courses, and short courses. Discourses related to these mostly turned towards career opportunities 'tailored for your success', 'mastering AI' and providing students the opportunity to 'learn from industry leaders'. Cumulatively, this AI content signalled that the university itself was oriented to AI innovation, was open to the business of developing AI for industry and for education, and was showcasing its academics as leading voices in AI research and policy.

Focus on Student AI Integrity and AI Adoption

In contrast to the overwhelmingly positive coverage of university AI research and development, student-focused AI content was primarily concerned with regulating student use of AI. The content here was largely informational and included explicit outlines of how students were to approach the use or prohibition of AI in university assessment. The ease with which these instructions were found—mostly as one of the top search results and nearly all on the first page of search results—demonstrated the centrality of AI-related integrity concerns. Some pages also featured content and resources for academics, including training in AI detection. The overall effect of this discourse was to centre the protection of degree integrity as a primary institutional concern.

A second lot of student-oriented content presented an 'AI-positive' approach, where access through a university commercial licence to AI products like Microsoft Copilot was outlined and promoted to students. The emphasis here was on the enhanced learning experience for students and the opportunities provided using new AI technologies. Some pages were also skills-focused, where students were directed to information or self-directed short courses where they could learn basic AI skills like prompt engineering. The discourses in these student-facing materials reveal an implicit tension: universities seek to portray AI use as sufficiently restricted so as to protect the integrity of their degrees, while at the same time promoting access to AI's affordances as essential for students to develop future-ready professional skills.

The discourses across both these thematic areas of AI technological embrace and student facing AI integrity were, at institutional level, largely uncritical. Certainly, in relation to technological development and tech partnerships, the emphasis was on purported benefits and opportunities. What critical content there was mostly focused on student 'misuse' of AI, with some information detailing limitations of the technology, particularly the potential inaccuracy of generated results. This produced an overall sense of students being prepared to be effective users of AI, with access to the latest products, while having some understanding of the limitations of these products.

However, while these two dominant themes could be observed across all universities scanned, they mask an underlying disparity: universities do not have equal access to AI resources, with significant implications for students attending different universities and the professions into which they are graduating.

Stratified Access to AI Tools

Among Australia's top-tier universities, distinct patterns emerge in terms of how students' and academics' access to AI tools is portrayed. These elite institutions possess considerable resources to purchase "cutting edge" AI products and services, often showcasing them prominently on their websites. Rather than developing AI tools in-house, universities publicise their partnerships with large, global tech companies, capitalising off computational and cloud software they neither own nor directly manage but still derive significant value from. For example, UNSW

has partnered with OpenAI to embed GenAI tool ChatGPT Edu into "the fabric of UNSW" (Hollingworth, 2024). The well-resourced, elite universities tended to highlight a broader range of these tools, in contrast to smaller, regional, and less resourced universities, which were more likely to feature just one or two, or none. These disparities highlight not just material inequalities but also the ways universities use the showcasing of AI to elevate their prestige and position themselves competitively in the global higher education market.

Large, well-resourced universities have become the engine rooms of data production and innovation in the EdTech sector, generating the research outputs, pedagogical content, and user data that fuel the industry's growth and profitability (Sadowski, 2019). At the same time, their greater financial resources enable these universities to be more selective in how they engage in the EdTech economy. They have greater say, for example, in determining the extent to which institutional data is made available for commercial use. This form of "digital enclosure" occurs when the data generated through use of AI tools are withdrawn from broader commercial access and "made more exclusive to the service providers whose operations generate the data" (Kuerbis & Mueller, 2023, p.1). Yet digital enclosure is a luxury afforded to only those universities able to pay for it. While Australia's top-tier universities can enhance their competitive advantage, leveraging their financial resources to access a broad array of commercial AI tools while controlling, to some degree at least, the collection and use of university-produced data, less advantaged universities often lack the means to opt out of data ownership and control arrangements misaligned with their own commercial and institutional interests.

The stratification we observed between top- and bottom-tier universities is not surprising. The ability to innovate with AI, secure access to cutting-edge tools while retaining a degree of data control, and strategically navigate AI-related risks and opportunities for competitive advantage mirrors long-standing patterns of privilege and institutional inequality. The danger is that these divides will become further entrenched over time, as AI becomes an unquestioned feature of the modern university. We are already beginning to see the solidification of a two-tiered system, with elite universities not only enjoying greater access to AI tools, but importantly also exerting more influence over the terms of that access, while less-resourced institutions are left scrambling to keep pace, reliant on generic, off-the-shelf tools with less say in their design or conditions of use.

Engagement and Thought Leadership

Describing scenarios for the future of work, Brown (2024, p.6) posited, "we need to ask if digital transformations are being used to increase the demand for knowledge and skills at all levels of the occupational structure, or leading to increasing segmentation and re-stratification of knowledge work?" Based on current evidence, the trajectory within universities at a structural level appears to align more closely with the latter. In this regard, there is a clear policy imperative to ensure that all students have access to AI tools, rather than limiting such opportunities to those at elite universities that can currently afford to provide them.

Even more fundamentally troubling, we argue, is the subordination of the social element of this socio-technical system. Based on evidence from our scan, in most instances the "social" dimension of AI for students has been reduced almost entirely to issues of ethics framed through academic integrity. This is not to suggest that a social focus is entirely absent—individual academics and specific courses do engage with this—but at the institutional level, the social remains secondary. That is, public-facing content on AI in universities casts the university itself as a technocratic institution rather than one committed to fostering critical reflection and advancing the public good.

This produces an impoverished view of AI in the social world. It leaves unexplored the social construction of AI, including the social harms reinscribed through the reproduction of social power relations, for instance in algorithmic decision-making that amplifies preexisting data biases against particular population groups. This was well documented in the Australian case of Centrelink's automated Online Compliance Intervention system, or "Robodebt". Universities could be leading this discussion about AI harms and technology's rightful place in education, but instead appear, on the surface at least, to be more concerned with promoting their own innovation credentials. This in turn has implications for the professionals students become and the professions they enter into. As Susskind and Susskind (2022) note, digital tools are frequently used to 'systematise past practices' rather than transform or innovate in the profession itself. If universities only emphasis is on AI's practical affordances, while students may learn to be adapt AI adopters and technocratic operators, they won't develop the critically evaluative sensibility needed to assess AI's broader social, ethical, and political implications. This denies them both important professional capabilities and a wider social imaginary beyond immediate careerist concerns (Harper & Saltmarsh, 2024). This also robs their future professions of critically evaluative thinkers (Bearman et al., 2023), important for students of all disciplinary background but particularly crucial for the community, social, and policy sectors, which are precisely the fields tasked with responding to and shaping the societal effects of AI (Henman, 2020).

Possibilities for Social

A rich social engagement and critique with AI is certainly possible, and there already exist some strong foundations. Harper and Saltmarsh (2024, pp.8–9) have called for a new social contract between higher education and society that re-centres the public good in practical ways directly connected to community and led by social science, humanities, and the liberal arts. Sociologists, critical policy makers, and other social scientists have a central role both in describing how different inequalities are shaped and reproduced through algorithmic technologies—both in organisations as well as on a wider global scale—and in using socially informed critique, interdisciplinary collaboration, and policy engagement (including refusal) to influence these dynamics (Zajko, 2022). Critical policy studies challenge the dominance of techno-rationalist practice in policy analysis (Griggs et al., 2001, p.1), including a focus on methods of AI decision-making and on wider

structural inequalities through which the technologies of governance are enacted (Paul, 2022). These all offer key principles for a socially critical approach to AI.

Beyond this are critical frameworks such as those described by sociologist Simon Lindgren (2023, pp.2–4). Drawing from Agger (2013), he outlines a social constructivist approach to AI rather than a positivist one, directing attention not only to historical and contextual factors but also to how power relations are structured and reproduced through technology. He calls for academics to cultivate a critical consciousness of AI's distortions, exploitation, discrimination, and other injustices in current and future society, and to develop approaches for addressing them. This involves critically examining AI in relation to the social, political, economic, and cultural contexts in which it is developed and applied, with the goal of moving beyond deterministic assumptions and generating knowledge of structural conditions that can foster meaningful social change. Adopting and promoting this type of thinking, we argue, is every bit as important as technical innovation and digital proficiency.

Conclusion

Artificial Intelligence (AI) is best understood as a sociotechnical system—shaped not only by algorithms and computational design, but also by the social, political, and economic structures in which it is developed, deployed, and contested. However, as our analysis shows, in the outward face of universities, it is the technical that is dominant. Given the neoliberal context of universities (Bok, 2009; Collini, 2012; Connell, 2019; Marginson, 2013), the dominant foregrounding of AI technical innovation on university websites can be readily understood as showcasing universities to their markets—including high fee-paying students—and the university's superior position. Thus, rather than 'revolutionising learning' or producing a break with past practice, the way in which universities are currently adopting and adapting to AI reflects pre-existing market logics and risks producing a technically adept, but professionally anaemic generation of graduates who have not been exposed to critical social debates around the adoption of AI.

The danger at this moment is that the very critical consciousness universities are meant to cultivate among future generations is being overshadowed by an emphasis on technical proficiency and marketable skills, leaving students ill-prepared to question, challenge, or shape the social world beyond the university. In the age of AI, the social sciences, including community and policy studies, arts, and humanities are uniquely positioned to reclaim a central role in the university by critically interrogating the ethical, cultural, and political dimensions of algorithmic systems. Rather than merely adapting to technological shifts, these disciplines can lead in shaping a new social contract for higher education—one that foregrounds human values, democratic accountability, and epistemic justice in the governance and deployment of AI and can imagine the good digital society (Mankoo & Digby, 2024). By embedding critical inquiry into AI's societal impacts, social science can offer essential frameworks for understanding not only what AI does, but also what it ought to do, thereby reaffirming its indispensable role in steering universities

toward socially responsive futures. By foregrounding questions of power, inequality, and representation, we can help ensure that AI serves public values rather than reproducing existing hierarchies.

References

Academy of the Social Sciences in Australia **(**2021, November 22). *State of the social sciences 2021*. Canberra: Academy of the Social Sciences in Australia. https://stateofthe socialsciences.org.au/wp-content/uploads/sites/4/2021/11/State-of-the-Social-Sciences-2021.pdf

Agger, B. (2013). *A critical theory of public life: Knowledge, discourse and politics in an age of decline*. Routledge.

Australian Financial Review (2024). *Best Universities Ranking*. Retrieved February 18, 2025. https://www.afr.com/lists-and-awards/best-universities-ranking

Australian Government (2022). *Finance 2022: Financial reports of higher education providers*. Australian Government, Department of Education. https://www.education.gov.au/download/17857/finance-2022-financial-reports-higher-education-providers/37982/2022-university-finance-summary-information/pdf

Ball, S. J. (2012). Performativity, commodification and commitment: An I-spy guide to the neoliberal university. *British Journal of Educational Studies*, *60*(1), 17–28.

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160–1173.

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: A critical literature review. *Higher Education*, *86*(2), 369–385.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code.* Polity.

Bok, D. (2009). *Universities in the marketplace: The commercialization of higher education*. Princeton University Press.

Brown, P. (2024). Education, opportunity and the future of work in the fourth industrial revolution. *British Journal of Sociology of Education*, *45*, 1–19. https://doi.org/10.1080/01425692.2023.2299970.

Burton, S., & Bowman, B. (2022). The academic precariat: Understanding life and labour in the neoliberal academy. *British Journal of Sociology of Education*, *43*(4), 497–512. https://doi.org/10.1080/01425692.2022.2076387

Collini, S. (2012). *What are universities for?* Penguin UK.

Collini, S. (2017). *Speaking of universities*. Verso Books.

Connell, R. (2019). *The Good University: What universities actually do and why it's time for radical change*. Bloomsbury Publishing.

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*, 50(7), 1087–1097. https://doi.org/10.1080/02602938.2025.2503964

Crawford, K. (2021). *The Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Croucher, G. (2023). Three decades of change in Australia's university workforce. *Melbourne Centre for the Study of Higher Education*. https://doi.org/10.26188/23995704

Australian Government Department of Education. (2018). International student data 2018. Australian Government Department of Education. https://internationaleducation.gov.au/research/international-student-data/Pages/InternationalStudentData2018.aspx

Department of Education. (2024a). *Selected higher education statistics – 2022 student data*. https://www.education.gov.au/higher-education-statistics/studentdata/selected-higher-education-statistics-2022-student-data

Department of Education. (2024b). *Australian universities accord final report document*. https://www.education.gov.au/australian-universities-accord/resources/finalreport

Dotan, R., Parker, L. S., & Radzilowicz, J. (2024, June). Responsible adoption of generative AI in higher education: Developing a "points to consider" approach based on faculty perspectives. In *Proceedings of the 2024 ACM conference on fairness accountability, and transparency,* Rio de Janeiro, Brazil, 2033–2046. https://doi.org/10.1145/3630106.3659023

Driessens, O., & Pischetola, M. (2024). Danish university policies on generative AI: Problems, assumptions and sustainability blind spots. *MedieKultur: Journal of Media and Communication Research*, *40*(76), 31–52. https://doi.org/10.7146/mk.v40i76.143595

EduRank. (2025). *Best Artificial Intelligence universities in Australia [2024 rankings]*. Retrieved June 13, 2025. https://edurank.org/cs/ai/au/

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Griggs, S., Mathur, N., & Jas, P. (2007). Editorial. *Critical Policy Studies, 1*(1), 1. https://doi.org/10.1080/19460171.2007.9518505.

Hannan, E., & Liu, S. (2023). AI: New source of competitiveness in higher education. *Competitiveness Review, 33*(2), 265–279. https://doi.org/10.1108/CR-03-2021-0045

Harper, J., & Saltmarsh, J. (2024). Pausing in pursuit of a new social contract between higher education and society. *Change: The Magazine of Higher Learning*, *56*(2), 4–10. https://doi.org/10.1080/00091383.2024.2322907

Henderson, M., Selwyn, N., & Aston, R. (2017). What works and why? Student perceptions of 'useful' digital technology in university teaching and learning. *Studies in Higher Education*, *42*(8), 1567–1579. https://doi.org/10.1080/03075079.2015.1007946

Henman, P. (2020). Improving public services using artificial intelligence: Possibilities, pitfalls, governance. *Asia Pacific Journal of Public Administration*, *42*(4), 209–221. https://doi.org/10.1080/23276665.2020.1816188

Hollingworth, D. (2024, December 17). *UNSW signs partnership agreement with OpenAI. Cyber Daily*. https://www.cyberdaily.au/tech/11507-unsw-signs-partnership-agreement-with-openai

Jensen, L. X., Buhl, A., Sharma, A., & Bearman, M. (2024). Generative AI and higher education: A review of claims from the first months of ChatGPT. *Higher Education*, 89, 1145–1161. https://doi.org/10.1007/s10734-024-01265-3

Kim, I., & Kuljis, J. (2010). Applying content analysis to web-based content. *Journal of Computing and Information Technology*, *18*(4), 369–375.

Komljenovic, J., Birch, K., Sellar, S., Bergviken Rensfeldt, A., Deville, J., Eaton, C., … & Williamson, B. (2025). Digitalised higher education: Key developments, questions, and concerns. *Discourse: Studies in the Cultural Politics of Education*, *46*(2), 276–292. https://doi.org/10.1080/01596306.2024.2408397

Kuerbis, B., & Mueller, M. (2023). Exploring the role of data enclosure in the digital political economy. *Telecommunications Policy*, *47*(8), 102599.

Lee, C. C., & Low, M. Y. H. (2024). Using GenAI in education: The case for critical thinking. *Frontiers in Artificial Intelligence*, *7*, 1452131.

Lindgren, S. (2023). Introducing critical studies of artificial intelligence. In S. Lindgren (Ed.), *Handbook of critical studies of artificial intelligence* (pp. 1–19). Edward Elgar Publishing.

Lodge, J. M. (2024). The evolving risk to academic integrity posed by generative artificial intelligence: Options for immediate action. *Tertiary Education Quality and Standards Agency*, pp. 2024–3008.

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, *49*(5), 651–664, https://doi.org/10.1080/02602938.2024.2309963

Mankoo, A., & Digby, J. (2024, August). *What are the possibilities of a good digital society? Introduction to a set of papers commissioned by the British Academy*. The British Academy. https://www.thebritishacademy.ac.uk/publications/introduction-to-discussion-paper-series/

Marginson, S. (2013). Global field and global imagining: Bourdieu and worldwide higher education. In M. David & R. Naidoo (Eds.), *The sociology of higher education* (pp. 19–35). Routledge.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. Illustrated ed. New York University Press.

OECD. (2023). Education at a glance. https://www.oecd.org/en/publications/education-at-a-glance-2023_e13bef63en.html

Paul, R. (2022). Can critical policy studies outsmart AI? Research agenda on artificial intelligence technologies and public policy. *Critical Policy Studies*, *16*(4), 497–509. https://doi.org/10.1080/19460171.2022.2123018

Sadowski, J. (2019). When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, *6*(1). https://doi.org/10.1177/2053951718820549 (Original work published 2019)

Selwyn, N. (2014). *Digital technology and the contemporary university: Degrees of digitization*. Routledge.

Selwyn, N. (2015). Data entry: Towards the critical study of digital data and education. *Learning, Media and Technology*, *40*(1), 64–82. https://doi.org/10.1080/17439884.2014.921628

Selwyn, N. (2019). What's the problem with learning analytics? *Journal of Learning Analytics*, *6*(3), 11–19.

Selwyn, N. (2024). On the limits of artificial intelligence (AI) in education. *Nordisk tidsskrift for pedagogikk og kritikk*, *10*(1), 3–14.

Shao, Z., Yuan, S., & Wang, Y. (2020). Institutional collaboration and competition in artificial intelligence. *IEEE Access*, *8*, 69734–69741.

Susskind, R., & Susskind, D. (2022). *The future of the professions: How technology will transform the work of human experts*. Oxford University Press.

Times Higher Education (2025). *Best universities in Australia 2025*. Retrieved March 13, 2025. https://www.timeshighereducation.com/student/best-universities/best-universities-australia.

UniversityRankings.com.au (2024). *Top 10 Universities in Australia 2024*. Retrieved February 18, 2025. https://www.universityrankings.com.au/top-10/

Vicsek, L. (2021). Artificial intelligence and the future of work–lessons from the sociology of expectations. *International Journal of Sociology and Social Policy*, *41*(7/8), 842–861. https://doi.org/10.1108/IJSSP-05-2020-0174

Williamson, B. (2017). Educating silicon valley: Corporate education reform and the reproduction of the techno-economic revolution. *Review of Education, Pedagogy, and Cultural Studies*, *39*(3), 265–288. https://doi.org/10.1080/10714413.2017.1326274

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass*, *16*(3), e12962. https://doi.org/10.1111/soc4.12962

Zouhaier, S. (2023). The impact of artificial intelligence on higher education: An empirical study. *European Journal of Educational Sciences*, *10*(1), 17–33. https://doi.org/10.19044/ejes.v10no1a17

8 AI and Indigenous Knowledge Holders

When Culture Becomes Security: Cultural Security and Epistemic Sovereignty in the Sámi Context

Barbara Schabowska

Introduction

At first glance, one might assume a parallel between the structure and connections of neural networks, the foundation of modern artificial intelligence (AI), and the social structures of Indigenous communities. One could say that both systems generate knowledge in a similar way. In each case, it is linked to a web of connections, relations, nodes, and interdependencies. Neural networks, through deep learning, can capture complex relationships between input and output within non-linear dependencies. The knowledge thus generated relies on the relations within the network – it emerges through the capacity to capture intricate connections.

Superficially, Indigenous Knowledge (IK) operates in a similar way, grounded in the relations between community members and their surrounding ecosystem. Yet this is where the analogy ends, because what seems similar is, in fact, the source of a deeply rooted difference. The epistemology enabling the construction of neural networks rests on entirely different foundations than Indigenous epistemology. The types of relations existing within these structures are fundamentally different. In the case of IK, they extend beyond data, entering the domain of social ties, traditions, culture, and ecology.

Future scenarios may unfold faster than expected; therefore, the social changes currently taking place in our societies as a result of the expanding use of AI technologies must be recognised and named to examine their consequences and to develop ways of mitigating them. Indigenous societies, in particular, provide a compelling case through which to analyse AI and the disruption of the social. This chapter explores how IK articulates resistance to AI – aligning itself with social relationality, cultural diversity, and the safeguarding of cultural security.

Whether using neural networks, statistical methods, or symbolic approaches, AI reconfigures knowledge by eliminating its creators, users, and, in Indigenous societies, specific knowledge holders. These knowledge holders can never truly be replaced, yet the appearance of AI on the horizon – especially in some Indigenous contexts – has caused significant concern. Among them are the Sámi, the Indigenous people of a region that today stretches across the northern parts of Sweden, Norway, and Finland and the Kola Peninsula in Russia, to whom this chapter refers. Place-based knowledge communities (Johnson, 2012) have been

DOI: 10.4324/9781003627920-8
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

quicker than other social groups to recognise that, in the face of AI's rapid development, their rights and identity are under threat. AI's disruption of the social lies in its capacity to destabilise the structure of Indigenous communities through the disturbance of internal relationships and symbolic interventions at epistemic, ontological, and political levels, which collectively render it a threat to their security.

I will approach AI both as a material construct, a data architecture, and as a discursive figure that shapes systems of knowledge. Following Haraway's relational ontology, I treat AI not as a tool but as embedded within a web of interdependencies – an idea resonant with Indigenous epistemologies in which agency is distributed. At the same time, I situate the reflection on the AI–IK relationship within the domain of societal security, examining how the development of AI-based technologies influences cultural security, understood here within the framework of the Copenhagen School.

Artificial Intelligence and Society

Driven by the rapid evolution of AI, societies now find themselves confronting a paradox of progress, gaining efficiency while gradually losing control over their own internal dynamics. There is no return to the pre-AI era, and the speed of this technological revolution is staggering. Current social acceleration unfolds not without desynchronisation (Rosa & Trejo-Mathys, 2013). As Kulveit et al. (2025) argue, this mismatch may gradually undermine human decision-making systems, leading to a loss of control over the economy, culture, and even the political order of nation-states.

The most recent attempts at systemic risk assessment provide what appears to be a comprehensive overview of the dimensions of these transformations, yet not of their consequences. As much as I appreciate the effort of Uuk et al. (2024) to identify 13 categories of systemic and societal risks associated with general-purpose AI, it is nonetheless disappointing that the mechanisms behind these risks remain unexplained. This concern applies particularly to the category of "Irreversible change," which refers to possible profound and long-term negative transformations in social structures, cultural norms, and human relationships. How such changes unfold, and what their broader societal implications might be, remains to be determined.

Some aspects of these transformations have been explored by Qian et al. (2024), who analyse the ethical, legal, and governance dimensions of AI's societal impact. Boretti (2024) offers a concise overview, highlighting the societal, and economic risks emerging from AI-driven quantum technologies, including issues of bias and fairness, transparency and explainability, job displacement, security, and ethical concerns, privacy issues, lack of regulation and standards, as well as the challenges posed by their exponential growth.

The mechanisms behind large-scale social change are complex and cannot be reduced to a single factor; rather, they result from the co-occurrence of multiple risks, though rarely all at once. Recent analyses of AI's impact on the social fabric reiterate and expand upon arguments articulated decades ago, long before

the emergence of the first ChatGPT. These include concerns about the potential misalignment between machine objectives and social values (Wiener, 1989), the absence of an ethical layer within machine rationality (Weizenbaum, 1994), and the inherently political and non-neutral character of technology (Winner, 1980).

These observations become particularly salient when examined through the lens of IK, which, by virtue of its dynamic and relational nature, poses a challenge to both digitalisation and AI (Ardalan, 2025; Laituri et al., 2023; Sletto et al., 2021).

AI undoubtedly presents certain opportunities for IK, particularly through its potential to preserve endangered languages and cultural expressions. Yet the risks associated with its application may outweigh these benefits. Especially in their impact on social relations. In both cases, however, the issue becomes one of cultural security (UNESCO, 1996), understood as the community's ability to preserve and develop its cultural identity, to protect the values of its symbolic culture, and to maintain the "dominant features that give culture its identity" (Czaja, 2008, p. 36).

In the context of Indigenous cultures, these features determine the distinctive nature of IK data, which encode the socio-ecological relations linking people, spirituality, and the natural environment.

Artificial Intelligence and Indigenous Knowledge

IK, as described by Díaz et al. (2015, p. 13), refers to "a cumulative body of knowledge, practice and belief, evolving by adaptive processes and handed down through generations by cultural transmission, about the relationship of living beings (including humans) with one another and with their environment." Like science, IK emerges within a specific social context; however, it is produced not in laboratories or universities but through everyday practices and interactions (Warren, 1991). Consequently, it evolves together with its environment.

A comparable capacity to evolve alongside its environment is often attributed to learning AI systems. Yet this adaptive quality is purely mechanical in nature, being dependent on data volume rather than on qualitative judgement. The calibration mechanism operates through the accumulation of quantitative inputs, whereby operational effectiveness is enhanced through scale, but never in the reverse direction – from qualitative reasoning to knowledge production. Consequently, AI outcomes are determined by the amount of data rather than by the quality or situated relevance of arguments. This mode of inference fails to account for social dynamics and contextual specificity, while simultaneously aspiring to generalise limited quantifications into universal claims.

IK does not advance claims to universality beyond the narrow scope of its own cultural group. It could therefore be described as consisting of "particular truths" (Padilla, 2001), specific to a given socio-cultural community. This is largely due to the central anchoring of Indigenous societies in the land – their relationship with the territory, which becomes an integral part of the community itself (Atleo & Boron, 2022), creates a unique and enduring bond.

This feature, poetically described by Kimmerer (2020), is intertwined with reciprocity and kinship, relations that encompass humans, animals, the natural

environment, and the spiritual beings associated with particular places. Multifaceted, interspecies connections and the ontology they express lead to Burkhart's profound observation that Indigenous communities do not operate within the concept of autotelic values (2019). IK, which integrates the socio-ecological relationships between humans, animals, and nature, resists representation in forms that exclude its inherent relationality and contextuality. A striking example of this difficulty can be found in attempts to cartographically record situated knowledge related to Sámi reindeer husbandry.

In efforts to capture the socio-ecological relationships among herders, reindeer, and the landscape, participatory mapping methods have been developed to enable knowledge holders to retain control over both the map and its narrative. Traditionally, the Sámi oriented themselves through narrative-based place names rather than by creating or using conventional maps. Knowledge of Sámi reindeer husbandry integrates information about herders, reindeer, and nature; it is deeply embedded in the local and socio-ecological context, reflecting patterns of land use and their transformations (Johnsen, 2016).

The participatory topological mapping method proposed by Johnsen et al. (2024) engages both reindeer herders and researchers, offering potential for the co-production of knowledge. Its topological dimension captures relations, connections, and spatial interdependencies collected through herders' narratives. Although the method is not without limitations, as the relevance and accuracy of such maps may quickly diminish, Johnsen et al.'s conclusion that "relations can tell different stories than locations of land use and land use change" (2024, p. 225) points directly to a key cultural feature of the Sámi world: relationality.

Preserving this relational dimension required not only participatory mapping but also control over validation and the scope of data disclosure. Interestingly, the Sámi herders themselves chose not to publish specific place names, or toponyms, explaining this decision by a concern that they might be misunderstood by users unfamiliar with Sámi cultural contexts.

Toponyms themselves are one of the traditional tools for transmitting knowledge: they assign unique names to places, binding the spiritual with the material and anchoring meaning in lived human experience. They are, in essence, emblems of relationality within Sámi society and culture, closely intertwined with oral history passed down through generations. The decision to withhold them, even under self-governed conditions, marks a fear not of visibility, but of epistemic displacement through interpretation, a concern deeply rooted in the awareness that the very act of translating oral, place-based traditions into cartographic form transforms the epistemic nature of what is being represented.

This reflects the broader Sámi understanding of knowledge as situated, relational, and embodied. Each body of knowledge has its own custodians, responsible for its intergenerational transmission and ethical use. AI's growing involvement, capable of "extracting" or reproducing such knowledge without cultural mediation, threatens these relations of guardianship and weakens the epistemic integrity of Sámi knowledge systems. In doing so, it contributes to a subtle reconfiguration of socio-ecological relations within these communities. The translation of embodied,

relational knowledge into data objects further displaces traditional authority, unsettling custodianship practices and eroding epistemic sovereignty.

AI-generated knowledge echoes the epistemic logic Haraway described as the "god trick of seeing everything from nowhere" (1988, p. 581): it claims objectivity by erasing its own location, mediation, and accountability. By contrast, objective vision is grounded in limited location and situated knowledge (Haraway, 1988, p. 583), rather than in decontextualised logics. As Lewis et al. (2024) observe, AI does not have an ethics problem; it has an epistemology problem. Indeed, the status of knowledge and the very notion of knowing in Indigenous communities differs fundamentally from the scientific epistemology that underlies algorithms and neural networks.

Classical accounts of scientific knowledge, to invoke Popper's metaphor, describe it as a building erected on piles, above a swamp (Popper, 2002). In this construction, what matters is not whether the piles reach any ultimate ground, but whether they are sufficiently stable to sustain the structure. The criterion of validity is thus pragmatic rather than ontological, and stability is procedural rather than rooted in any given foundation. Scientific knowledge is not grounded in place, relationality, or tradition; rather, it deliberately rises above them. It aspires to universality and operates through abstraction and, following Kuhn (1962), is organised around paradigms that define what counts as legitimate knowledge.

This epistemic orientation can be traced back to the Enlightenment commitment to reason as a universal faculty, capable of grounding knowledge independently of tradition or place (Kant, 1991) and characterised by an emphasis on progress and procedural legitimacy. AI emerges directly from this intellectual lineage. Detached from relational grounding and contextual accountability, AI values functional performance over situated meaning.

IK, by contrast, recognises its swamp. It is not bracketed out of the epistemic structure but constitutes an integral part of it. Its epistemic authority is situated, relational, and context-dependent, rather than universalist or paradigm-driven.

The limits arising from these epistemic differences inevitably lead to distortion and misrepresentation: precisely what the reindeer herders feared. Such distortions stem not only from epistemological asymmetry but from the very act of translation – the transposition of one language (IK) into another (AI). As Quine reminds us, meaning is never fully determined by empirical data (2013); radical translation is, by definition, impossible, and differences – and thus limitations– remain irreducible. The gap between algorithmic language and Indigenous systems of meaning is, by nature, unbridgeable, shaping both the comprehension and reception of IK once digitised. This, as Quine would insist, makes misrepresentation not accidental but structural. Such epistemic dislocations may take a material form, manifesting in false outputs or misleading inferences generated by large or small language models. Or, over time, a symbolic one. Distorted narratives about Indigenous peoples, stripped of authorship, deprive them of the power to tell their own stories. As Said (1993) observed, losing the power to narrate is itself a form of colonisation.

Questions of data decolonisation, and ethical responsibility increasingly frame academic debates on AI and IK (Rana, 2025; Belarde-Lewis et al., 2024; Roberts & Montoya, 2023). The distortion of data and the consequent production of false representations of Indigenous communities is, as also noted by Perera et al. (2025) after their decade-long review of relevant literature up to January 2023, among the most pressing concerns. It is regrettable, however, that while identifying areas for future research, Perera et al. fail to address the social consequences of these phenomena, such as identity, economic, and security implications for Indigenous communities. Since the discussion primarily concentrates on the methodological preconditions for constructive collaboration between AI and IK, it tends to drift away from the social sciences, thereby avoiding a critical examination of AI's societal consequences and responsibilities in the broader process of transformation is referred to here as the disruption of the social.

Data distortion frequently manifests as the reduction of Indigenous identities to rigid classificatory schemes, producing misrepresentation, and epistemic erasure (Walter et al., 2021). It may also have economic consequences, as AI-driven policy systems reproduce inaccurate profiles of Indigenous needs, affecting resource distribution and development strategies. In security contexts, data distortion contributes to the algorithmic framing of Indigenous communities as objects of risk or surveillance, transforming cultural practices into security concerns (Perera et al., 2025; Crawford, 2021).

Technology accelerates transformations in ontology and social relations, including those that extend beyond human connections to encompass interspecies kinship, as in the case of Sámi communities. AI, as one of the actors co-constituting the contemporary human world, reshapes existing hierarchies of agency, a process insightfully recognised by Haraway (2003, 2016). The example of mapping, when algorithmically multiplied and transposed into the domain of AI, demonstrates how technology may sever intergenerational transmission, unsettle collective decision-making, and estrange kinship with non-human entities translating them out of the language of relation and, thereby distorting the very fabric of the social world. From the perspective of societal security, AI represents an unprecedented threat to the continuity and diversity of social relations, introducing profound tension between technology and culture.

Artificial Intelligence and Cultural Security

A notable lack of reflection on cultural security appears in several recent works on AI and cultural heritage, including the otherwise impressive volume by Ardalan et al. (2025). In that collection, security is addressed solely in terms of data protection, despite the stated aim of exploring how AI contributes to preserving culture and tradition, a matter inseparable from cultural security. This conceptual gap is even clearer in Linkola-Aikio et al. (2025), where the focus on Sámi art, culture, and digital heritage overlooks the implications of applying machine and deep learning to cultural expression. Even the most rigorous analysis of AI's potential to safeguard or promote heritage cannot substitute for such reflection.

The persistent absence of cultural security from analyses of societal security results from the tendency to treat culture not as a fundamental element but as an ancillary dimension, rarely acknowledged as a key factor of sustainable development. This becomes evident when examining The 2030 Agenda for Sustainable Development (United Nations, 2015). None of its 17 goals is devoted specifically to culture; it appears only as a secondary aspect within several seemingly minor targets. For instance, Target 11, addressing Sustainable Cities and Communities, in subsection 11.4, calls to "strengthen efforts to protect and safeguard the world's cultural and natural heritage." Yet the notion of sustainable culture is entirely absent from the document.

Data-driven classification does not merely reflect social realities but actively reshapes them, producing distorted representations that carry material and institutional consequences. Eubanks (2018) demonstrates how algorithmic welfare systems translate biased data and rigid classificatory logics into administrative decisions that systematically misrepresent marginalised communities and reproduce structural inequality.

It is precisely this marginal positioning of culture within dominant frameworks of societal security that allows AI-driven interventions to appear neutral or technical, even as they reorganise cultural practices and relations in ways that generate cultural insecurity.

From a post-humanist perspective, the relational ontology characteristic of Indigenous cosmologies – open to interspecies kinship and resonant with Haraway's idea of companion species – could, at least theoretically, support the inclusion of technology within the framework of social relations. In such a view, AI would not necessarily constitute a threat to the Sámi but could become a co-constitutive element of knowledge, which has always had multiple authors. After all, the Sámi sustain relations with beings that are material, spiritual, and symbolic alike – so why not with AI? Because AI, as recent developments suggest, has become a source of cultural insecurity, evidenced by its growing securitisation by Indigenous communities themselves, who perceive it as an existential threat.

AI Securitisation

Securitisation theory, developed by the Copenhagen School, is one of the leading contemporary approaches in international relations, explaining how ordinary policy matters become transformed into security issues (Wertman & Kaunert, 2022). The concept of security as extending beyond the military domain to include culture, alongside language, religion, national identity, and customs, as a key component of societal security (Buzan, 1991), provides a framework for developing the notion of cultural security. Through this lens, the societal consequences of AI's expansion within Indigenous communities can be more precisely understood.

The mechanism of securitisation is rooted in the performative function of language naming a phenomenon as threatening to a given audience effectively makes it so. The theory thus perceives security and threats as social constructs, produced through speech acts and practices. The threat itself may be real or imagined – what matters is that a community recognises it as such under the influence of discourses disseminated by key social and political actors.

Securitisation legitimises exceptional measures undertaken by the state (Buzan et al., 1998) and functions as a driver of politicisation, the process through which an issue or practice becomes subject to political debate, decision-making, or struggles overpower (de Wilde et al., 2016). In recent years, this framework has proved particularly effective in explaining how emerging technologies, including AI, are reframed as matters of security and governance rather than merely as technical instruments. A model case of AI securitisation can be observed within the Sámi community. In April 2024, the Sámi Council adopted the Roadmap of Sámi Ownership and Data Access (SODA) principles (Saami Council, 2024). SODA outlines a set of guidelines for handling Sámi data, grounded in the CARE framework, the principles of Collective Benefit, Authority to Control, Responsibility, and Ethics. The roadmap was introduced to protect Sámi data, knowledge, and culture from unauthorised use, and it has since become central to debates on data sovereignty in the age of AI (Fjellheim, 2024). This marks a distinct moment in which AI is articulated not merely as a technological innovation but as a potential threat to cultural identity and epistemic sovereignty, thereby activating the logic of securitisation within an Indigenous context.

The SODA principles are directly referenced in a document prepared by the Expert Mechanism on the Rights of Indigenous Peoples (EMRIP, 2025, p. 13) under the mandate of the United Nations Human Rights Council. The right to data, including their collection and disaggregation, implies the right to self-governance, as these data concern, among other things, language and culture, and thus constitute "a cultural, strategic and economic resource for Indigenous Peoples" (EMRIP, 2025, p. 4). Consequently, the protection of data fundamental to cultural identity and integrity must be recognised as a matter of security (UNESCO, 2001). Such data represent strategic information about a community – the very core of its security (Nye, 2011; Castells, 2010; Held, 1995), a point that the everyday experience of the disinformation age makes abundantly clear. The Sámi case thus illustrates how the politics of data intersects with broader questions of cultural security, positioning Indigenous data governance as both a political and existential safeguard.

Building on the earlier discussion of securitisation as a response to AI-driven cultural risk, the Sámi case demonstrates how existential concerns over culture and sovereignty are reframed as matters of security through data governance. This reframing has translated into concrete policy action: in October 2025, the Nordic ministers for culture funded a Nordic Network for Language Models for Artificial Intelligence under the Nordic-Baltic Centre for AI (Nordic Council of Ministers, 2025) to counter the algorithmic marginalisation of small cultures and languages. Here, securitisation operates as a strategy of agency, converting existential vulnerability into institutional mechanisms aimed at protecting cultural diversity, epistemic sovereignty, and the integrity of meaning in AI-mediated language systems.

At the meeting of the Nordic ministers for culture, where the decision was made to fund a new Nordic Network for Language Models for Artificial Intelligence, the discussion reportedly focused on the impact of AI on the Nordic cultural landscape, particularly in relation to copyright and creative industries. This framing, rooted

in a policy discourse that prioritises economic risk management over social risk considerations, however, remains notably narrow, overlooking the broader social consequences that accompany cultural transformation under technological acceleration. The tendency to justify such initiatives through the lens of intellectual property law reflects a persistent blindness within European policymaking to the importance of cultural security as a key dimension of social policy.

Indeed, the question of intellectual property and authorship poses a particular challenge in Indigenous contexts, where knowledge is collectively produced and transmitted across generations. Yet it is not ownership itself that stands most at risk in the face of AI's unregulated extraction of Indigenous data. It is the cultural security of those whose knowledge that is being appropriated. The issue, thus, extends beyond copyright; it reaches into the foundations of cultural identity and self-definition. Few policy debates acknowledge that the problem with AI lies not merely in intellectual property but in the gradual redefinition of cultural ownership, a shift from the language of law to that of security, and ultimately, to the politics of survival.

Artificial Intelligence, Cultural Security, and Mechanisms of Disruption

At first glance, neural networks and Indigenous social networks look oddly alike; both are built on relations, nodes, and flows that bind one element to another. Yet the semblance is superficial. In neural structures, relations are statistical dependencies meant to optimise prediction; in Indigenous worlds, they are dynamic, context based, and ecological, which generate obligations and meaning. I will argue that the difference here marks the boundary between technology and culture itself.

This tension produces not a single disruption but a pattern of overlapping fractures, that unfold simultaneously across epistemic, social, and political domains. In the context of IK and social transformation, these disruptions emerge through interconnected processes.

At the social and political levels, AI reconfigures structures of authority and responsibility. Communal guardianship and relational governance are gradually absorbed into system logics, while technological empowerment turns into dependence as sovereignty and decision-making are outsourced to code. Culture itself becomes framed as a security issue, intensifying social polarity and epistemic distrust.

Taken together, these mechanisms weaken social bonds, erode cultural resilience, and unsettle intergenerational continuity. Yet they also expose what AI cannot compute: relational accountability and collective meaning. What is at stake is not only IK but the cultural security of its holders and the integrity of the community itself. Yet, the Sámi case shows that disruption may also generate new forms of agency. The reframing of AI as a security issue and the integration of certain principles into governance enable these communities to transform resistance into a strategy of resilience.

The algorithm is becoming the new dogma, one that echoes the relentless rationality of the Enlightenment. Emancipation and progress have always carried their

inhuman face, as experienced not only by the Sámi, but also by European Jews and the Catholic Church. Likewise, AI-driven technologies may unfold as a new, posthuman form of order, threatening cultural and social security at multiple levels.

Yet within this tension between technological rationality and relational epistemology lies a real possibility. If AI can disrupt relations, it can also be designed to sustain them. A conscious policy of cultural security, understood not as a reaction to AI's securitisation but as a reinforcement of cultural continuity, could redefine the balance between knowledge and power. Indigenous worlds, in turn, challenge us to reconsider what intelligence is and how it relates to reason. Technology is never autonomous, it emerges through relations that entangle humans, non-humans, and matter itself. It is precisely this relationality that, despite everything, calls for culture to act as a mediating platform in this social transformation. Culture is relation itself.

References

Ardalan, I. D. (2025). Embedding cultural heritage into AI. In I. D. Ardalan, A. Banifatemi, F. Gonzalez, M. Ingram, R. Moradinezhad, & L. Williams (Eds.), *AI for community: Preserving culture and tradition* (pp. 1–30). CRC Press.

Ardalan, I. D., Banifatemi, A., Gonzalez, F., Ingram, M., Moradinezhad, R., & Williams, L. (Eds.). (2025). *AI for community: Preserving culture and tradition* (1st ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781003517115

Atleo, C., & Boron, J. (2022). Land is life: Indigenous relationships to territory and navigating settler colonial property regimes in Canada. *Land, 11*(5), 609. https://doi.org/10.3390/land11050609

Belarde-Lewis, M., Littletree, S., Braine, I. R., Srader, K., Guerrero, N., & Palmer, C. L. (2024). Centering relationality and CARE for stewardship of Indigenous research data. *Data Science Journal, 23*(1), 32. https://doi.org/10.5334/dsj-2024-032

Boretti, A. (2024). Technical, economic, and societal risks in the progress of artificial intelligence-driven quantum technologies. *Discover Artificial Intelligence*, *4*(1), 67. https://doi.org/10.1007/s44163-024-00171-y

Burkhart, B. (2019). *Indigenizing philosophy through the land: A trickster methodology for decolonizing environmental ethics and indigenous futures.* Michigan State University Press.

Buzan, B. (1991). New patterns of global security in the twenty-first century. *International Affairs, 67*(3), 431–451. https://doi.org/10.2307/2621945

Buzan, B., Wæver, O., & de Wilde, J. (1998). *Security: A new framework for analysis*. Lynne Rienner Publishers.

Castells, M. (2010). *The power of identity* (2nd ed., Vol. 2). Wiley–Blackwell.

Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Czaja, J. (2008). *Kulturowe czynniki bezpieczeństwa*. Krakowskie Towarzystwo Edukacyjne – Oficyna Wydawnicza AFM.

de Wilde, P., Leupold, A., & Schmidtke, H. (2016). Introduction: The differentiated politicisation of European governance. *West European Politics, 39*(1), 3–22. https://doi.org/10.1080/01402382.2015.1081505

Díaz, S., Demissew, S., Carabias, J., et al. (2015). The IPBES conceptual framework: Connecting nature and people. *Current Opinion in Environmental Sustainability, 14*, 1–16. https://doi.org/10.1016/j.cosust.2015.04.001

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Expert Mechanism on the Rights of Indigenous Peoples. (2025). *Right of Indigenous Peoples to data, including with regard to data collection and disaggregation: Study and advice* (A/HRC/EMRIP/2025/2). United Nations Human Rights Council. https://undocs.org/A/HRC/EMRIP/2025/2

Fjellheim, R. (2024). Commentary on the SODA principles and their implications for Sámi data management. *Acta Borealia, 41*(2), 127–128. https://doi.org/10.1080/08003831.2024.2410114

Haraway, D. J. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599. https://doi.org/10.2307/3178066

Haraway, D. J. (2003). *The companion species manifesto: Dogs, people, and significant otherness*. Prickly Paradigm Press.

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene.* Duke University Press.

Held, D. (1995). *Democracy and the global order: From the modern state to cosmopolitan governance*. Polity Press.

Johnsen, K. I. (2016). Land-use conflicts between reindeer husbandry and mineral extraction in Finnmark, Norway: Contested rationalities and the politics of belonging. *Polar Geography*, *39*(1), 58–79. https://doi.org/10.1080/1088937X.2016.1156181

Johnsen, K. I., Westerveld, L., Persson, A.-M., Stenberg, J., & Juntti, K.-J. (2024). Participatory topological mapping: A novel approach for exploring and communicating situated knowledge of complex socio-ecological systems. *Methodological Innovations, 17*(4), 215–228. https://doi.org/10.1177/20597991241287113

Johnson, J. T. (2012). Place-based learning and knowing: Critical pedagogies grounded in Indigeneity. *GeoJournal, 77*, 829–836. https://doi.org/10.1007/s10708-010-9379-1

Kant, I. (1991). An answer to the question: "What is Enlightenment?" (H. B. Nisbet, Trans.). In H. S. Reiss (Ed.), *Kant: Political writings* (pp. 54–60). Cambridge University Press. (Original work published 1784)

Kimmerer, R. W. (2020). *Braiding sweetgrass: Indigenous wisdom, scientific knowledge, and the teachings of plants.* Penguin Books.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., & Duvenaud, D. (2025). *Gradual disempowerment: Systemic existential risks from incremental AI development. arXiv.* https://arxiv.org/abs/2501.16946

Laituri, M., Luizza, M. W., Hoover, J. D., et al. (2023). Questioning the practice of participation: Critical reflections on participatory mapping as a research tool. *Applied Geography, 152*, 102900. https://doi.org/10.1016/j.apgeog.2023.102900

Lewis, J. E., Whaanga, H., & Yolgörmez, C. (2025). Abundant intelligences: Placing AI within Indigenous knowledge frameworks. *AI & Society, 40*, 2141–2157. https://doi.org/10.1007/s00146-024-02099-4

Linkola-Aikio, I., Keskitalo, P., Ballardini, R., & Sarantou, M. (Eds.). (2025). *Digital indigenous cultural heritage.* Springer Nature Switzerland AG

Nordic Council of Ministers. (2025, October 29). *Nordic ministers of culture invest in joint network for AI language models*. https://www.norden.org/en/news/nordic-ministers-culture-invest-joint-network-ai-language-models

Nye, J. S., Jr. (2011). *The future of power*. PublicAffairs.

Padilla, A. M. (2001). Quantitative methods in multicultural education research. In J. A. Banks & C. A. McGee Banks (Eds.), *Handbook of research on multicultural education* (pp. 127–145). Jossey-Bass.

Perera, M., Vidanaarachchi, R., Chandrashekeran, S., Kennedy, M., Kennedy, B., & Halgamuge, S. (2025). Indigenous peoples and artificial intelligence: A systematic review and future directions. *Big Data & Society, 12*(2), 1–22. https://doi.org/10.1177/20539517251349170

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Original work published 1934)

Rana, V. (2025). Indigenous data sovereignty: A catalyst for ethical AI in business. *Business & Society, 64*(4), 635–640. https://doi.org/10.1177/00076503241271143

Roberts, J. S., & Montoya, L. N. (2023). *In consideration of Indigenous data sovereignty: Data mining as colonial practice.* arXiv:2309.10215. 10.48550/arXiv.2309.10215

Rosa, H., & Trejo-Mathys, J. (2013). *Social acceleration: A new theory of modernity*. Columbia University Press. https://doi.org/10.7312/rosa14834

Saami Council. (2024, April 28). *Sámi Ownership and Data Access principles (SODA).* https://static1.squarespace.com/static/5dfb35a66f00d54ab0729b75/t/66bc919c80f68e5932bfb470/1723634076584/SODA+eng.pdf

Said, E. W. (1993). *Culture and imperialism*. Alfred A. Knopf.

Sletto, B., de la Torre, G. B., Luguana, A. M.-L., & Pereira Júnior, D. (2021). *Walking, knowing, and the limits of the map: Performing participatory cartographies in Indigenous landscapes. Cultural Geographies, 28*(4), 697–715.

UNESCO. (1996). *Our creative diversity: Report of the World Commission on Culture and Development*. https://unesdoc.unesco.org/ark:/48223/pf0000104415

UNESCO. (2001). *Universal declaration on cultural diversity.* https://unesdoc.unesco.org/ark:/48223/pf0000128347

United Nations. (2015). *Transforming our world: The 2030 Agenda for Sustainable Development* (A/RES/70/1). United Nations. https://doi.org/10.18356/9789210056083

Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., Slattery, P., & Prunkl, C. (2024). *A taxonomy of systemic risks from general-purpose AI. arXiv.* https://arxiv.org/abs/2412.07780

Qian, Y., Siau, K. L., & Nah, F. F. (2024). Societal impacts of artificial intelligence: Ethical, legal, and governance issues. *Societal Impacts*, *100040.* https://doi.org/10.1016/j.socimp.2024.100040

Quine, W. V. O. (2013). *Word and object.* MIT Press. (Original work published 1960)

Walter, M., Lovett, R., Maher, B., Williamson, B., Prehn, J., Bodkin-Andrews, G., & Lee, V. (2021). Indigenous data sovereignty in the era of big data and artificial intelligence. *Australian Aboriginal Studies*, *2021*(2), 36–50. https://doi.org/10.1002/ajs4.141

Warren, D. M. (1991). *Using indigenous knowledge in agricultural development.* World Bank Discussion Paper No. 127. World Bank.

Weizenbaum, J. (1994). *Computer power and human reason: From judgment to calculation.* W. H. Freeman.

Wertman, O., & Kaunert, C. (2022). The audience role in securitisation theory. *Strategic Assessment: A Multidisciplinary Journal on National Security*, *25*(3), 67–81.

Wiener, N. (1989). *The human use of human beings: Cybernetics and society* (Rev. ed.; with an introduction by S. Heims). Free Association Books.

Winner, L. (1980). Do artifacts have politics? *Daedalus, 109*(1), 121–136.

9 Masculinism and Gender Binarism in AI

How Artificial Intelligence Reproduces Patriarchy in the Digital World

Bob Pease

Introduction

In the last 20 years, we have witnessed significant developments in artificial intelligence (AI) in general and machine learning as a subset of AI in particular (Verdegem 2022). Machine learning is a form of computational pattern finding that uses statistics to construct mathematical models to train computers to undertake classification tasks without explicit human instructions. It is premised upon the idea that the way to solve problems is to make something as effective as possible through the process of optimisation (McQuillan 2022). Optimisation requires that knowledge be datafied to reduce its complexity and to create the conditions necessary to formulate what is seen as a superior solution to the problem. Such developments are lauded by many and critically interrogated by others. I argue that this form of statistical reasoning obscures the processes that reinforce social inequalities, including those of gender and gender identity.

AI is a form of what McQuillan (2022) calls 'tech-solutionism'. This is an updated version of scientism, which is a belief that the scientific method is the only way to address problems that are structurally located but perpetuated by humans. To the extent that the computational classifications of AI replace human judgement and dialogue, they lead to a form of 'thoughtlessness' (McQuillan 2019). While, in theory, AI can be used to mitigate and challenge oppressive power relations, Varon and Pena (2021) argue that AI and digital technology are being used to objectify and exploit marginalised groups by imposing ways of thinking and being, and denying the existence of alternative forms of knowledge.

In this chapter, I examine the patriarchal ideologies and social practices embedded within the concept of dataism, and the algorithms and governance of AI, and their implications for gender and sexual discrimination, epistemic injustice, and gendered violence in the online world. I outline the limitations of technical solutions for addressing 'gender bias', greater diversity and equity within tech companies, and ethical frameworks that do not address the socio-political problems underpinning AI. I argue that only an ethic of care, in association with the practice of mutual aid and commoning, can address the concerns raised by feminist, queer, and decolonial critiques of AI.

DOI: 10.4324/9781003627920-9
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

AI, Dataism, and the View from Nowhere

Dataism is a term used to describe the belief that we can best understand the world, and address the problems within it, through the application of quantitative scientific methods that supposedly represent an objective rational perspective, a view from nowhere, where truth is propagated irrespective of context (Little & Winch 2021a). In representing knowledge as neutral and universal, when it is actually specific and contingent, it entrenches hierarchy and masculinism. Masculinism was originally defined by Brittan (1989) as the 'ideology of patriarchy', in that it involves an acceptance of the so-called naturalness of gender dimorphism and biological gender. Masculinism privileges and universalises all things coded as masculine and subordinates alternative ways of knowing and understanding (Nicholas et al. 2024).

As Little and Winch (2021a) remind us, data is not neutral, and when information is decontextualised from its situatedness and cultural context, it reinforces existing forms of inequality and exploitation. As such, dataism is an emerging ideology that marginalises alternative epistemologies and fuels the far-right (Varon & Pena 2021). For example, Richards (2020) identifies the role of 'big data' in the social media communications of racially motivated far-right groups in the United States. D'Ignazio and Klein (2020) refer to it as 'big dick data' to describe projects that are characterised by masculinism and an inflated use of scientific and technical capacities. Drawing upon Haraway, Blas et al. (2025) use the language of 'informatics', the tendency to view the world primarily as information and data, and 'informatics of domination', to describe the impact of patriarchal power on information technologies.

While non-Western cultures can also be sexist, racist, and homophobic, because AI is underpinned by Western colonial values and epistemologies, it tends to perpetuate sexist, racist, and homophobic content (Mollema 2024). Couldry and Mejias (2019) refer to the abstract quantification methods of computing as 'data colonialism', to draw attention to the power imbalances embedded in the ways that data is extracted. To develop a decolonial AI, it would be necessary to abolish the false universalism that drives AI (Couldry & Mejias 2019).[1]

All epistemological perspectives reflect different standpoints in relation to class, race, gender, and other social divisions (D'Ignazio & Klein 2023). Given who is in the room when AI is designed, it is inevitable that it is a privileged standpoint that will be reflected in the algorithms (McQuillan 2022). In addressing 'bias', then, the aim is not to strive for neutrality as an ideal (something that is impossible to achieve), but rather to situate different perspectives in their cultural context. In this sense, the sexism, gender binarism, racism, and colonialism embedded in AI technologies are not 'glitches' to be eliminated by better technical operations, but rather are an inevitable result of ideologies embedded within the technology that cause bodily harm (Basset 2023).

While AI is not inherently fascist, as it is premised upon the reduction of knowledge to exclusionary models of abstracted patterns, it shares the same assumptions as right-wing populism and fascism (McQuillan 2019). In both fascism and AI, gender inequality is premised upon innate biological differences. Consequently, AI can easily perpetuate fascism and patriarchy (Little & Winch 2021b).

Masculinism, Patriarchy and Transhumanism in the Tech Industry

It has been well noted that women are under-represented in the techno-industry (D'Ignszio & Klein 2020; Wajcman & Young 2023; Dai 2024). Only 22% of AI professionals globally are women, while 78% are men. Furthermore, only 1% of AI researchers were women in a 2016 study (Dai 2024). While wealthy white men continue to own and manage tech companies in the Global North, it is inevitable that masculinism and patriarchy will continue to be perpetuated in the industry (Gestoso 2019). Consolidating the tools of data science in the hands of a small privileged group of men means that they are able to obscure the politics underpinning the technologies (D'Ignazio & Klein 2020).[2]

Many years ago, feminists drew attention to the link between masculinity and technological rationality (Wajcman 1991). Masculinity is directly associated with technology and the disproportionate power that accrues to men and boys arising from that link. While women were involved in the early days of computing (Abbate 2012), men and boys claim that they have a special aptitude for technological knowledge and they derive a privileged masculine identity from their engagement with technology (Salter 2018). Salter identifies this as 'geek masculinity', an alternative pathway for men to attain power and status. Geek masculinity is associated with the design of new technologies, the rise of AI, and online abuse.

Little and Winch (2021a) define the billionaire founders of tech companies as 'the new patriarchs of digital capitalism', who perpetuate gendered hierarchies within a patriarchal network. Focusing on Elon Musk, Jeff Bezos, Mark Zuckerberg, Peter Thiel, and Larry Page, Little and Winch demonstrate the role they play in defining hegemonic geek masculinity. They also explore how women, such as Sheryl Sandberg, reconcile themselves to this misogynistic milieu by 'leaning in' to the male-dominated workplace. This leaning in is framed by Little and Winch (2021a) as a form of 'bargaining with patriarchy'.

Although women are increasingly participating in technology industries (Butcher et al. 2022), geek culture is still rooted in racism, misogyny and male supremacy. Lockhart (2015) argues that the belief in masculinity and gender as absolute and innate fuels the rise of the neo-reactionary movement among technology professionals in Silicon Valley, which opposes modernity and democracy in their support for technological progress. Neo-reaction contributes to the rise of transhumanism (Lockhart 2015).

Transhumanism is an intellectual and philosophical movement that uses technology and science to advocate for the transcendence of the limitations of the human condition. Mendz and Cook (2021) refer to transhumanism as a new form of eugenics in the search for transcendence of the human body and the elimination of bodily pain. As the technology develops, it starts to shape men's subjectivities and leads to a transhumanist masculinity (Keeling 2012). The aim is to create a 'new man' through technological innovation.

Male transhumanists are also concerned about creating a 'new woman', one that makes men feel good about themselves and can infer a man's desires (Wilshire 2024). Chatbots developed by men are often designed with feminine voices and

female names, such as Apple's Siri and Amazon's Alexa. They perform tasks associated with women and emulate subservient attitudes. Men, in particular, sometimes become emotionally attached to these technologies (Costa & Ribas 2019). The representation of AI femininity that caters to patriarchal heterosexual norms is evident in many popular films, including The Stepford Wives (2004), Ex Machina (2014), and Her (2023). The technological representation of femininity in such films reinforces gendered cultural codes, whereby women are envisioned to fulfil male fantasies in terms of household and sexual roles to constitute the 'perfect woman', as envisioned by male creators (Dai 2024).

AI and the Problem of the 'Problem of Bias'

AI, as noted, is not gender or value neutral. Evidence of gender bias in AI is frequently cited (Schopmans & Cupac 2021; Ricaurte 2022; Khan 2024), in that traditional gender norms are deeply embedded in the technology. It is a product of the cultural context, and it reflects its creators' values (Aavik et al. 2024). It is also based upon statistical patterns derived from literature that is not up-to-date or informed by critical theory. In a study undertaken in 2024, the latest source of literature on questions posed about what is a man and what is masculinity were published in 2012, and most of the references cited were published before 2000 (Walther et al. 2024).

Some explanation of 'gender bias' is given by the gendered basis of the production and consumption of ChatGPT and other AI technologies. Research demonstrates that men constitute two-thirds of users of ChatGPT (Khan 2024). Also, given the wider gender-unequal world in which AI is located, it is not surprising that this technology would mirror the gender bias in the wider society. When AI is used as a recruiting tool, it consistently ranks men above women (Gestoso 2019), as was demonstrated in 2015 when it was discovered that Amazon's hiring algorithm always recommended men over women applicants (Wilshire 2024).

Gender bias embedded in algorithms and data sets results in part from the intentionality of the software developers who are predominantly wealthy, white, heterosexual cisgender men in the Global North (Dwiputri n.d.). The underrepresentation of women in science, technology, engineering and maths (STEM) is widely reported. A World Economic Forum Report noted in 2023 that women were only 29% of this labour force (UN Women 2024).

As noted, concerns and criticisms of AI are often framed in terms of bias. It is said that AI is biased against women, nonbinary people and people of colour. In this context, the desired aim is to produce an algorithm without prejudice or bias (Drage & Frabetti 2023). However, Crawford (2021) argues that the aim to produce 'fairer systems' through mathematical parity does not contend with the social, political and economic structures that underpin AI. Narrow technical solutions focused on statistical bias do not address the structural inequalities, either within AI or society in general.

Similarly, concerns with individual AI developers and their motivations reflect the limitations of liberal anti-discrimination discourses that focus on disadvantage

(Hoffman 2019). Developer bias is only one form of AI-based injustice (Miragoli 2024). Less attention is given to the normative conditions that produce and perpetuate the interests of advantaged subjects and the various ways in which structural processes reproduce social hierarchies. Miragoli (2024) refers to this as 'epistemic conformism' embedded within AI technology that treats information as relevant solely through it being statistically dominant. This discrimination embedded within computer code is what Noble (2018) identifies as 'algorithms of oppression'. Such discrimination also goes beyond epistemic bias, however, as it perpetuates structural inequalities in gendered labour, through exacerbating job insecurity, especially for women (Mohla et al. 2024).

Lindgren (2024) also argues that we should move beyond the narrow focus on 'bias' and focus more on algorithmic oppression, dispossession and exploitation. We cannot address 'bias' within the algorithms if the humans who designed them will not accept the so-called 'debiasing'. Furthermore, sexism, transphobia and racism are not simply 'biases' that can be corrected, as they are deeply embedded in history, language, ideology and social structures which cannot be addressed solely through a technological fix. The algorithms reflect an unequal and violent gender and racial order.

Drage and Frabetti (2023) are concerned about the impact of AI on people's bodies, subjectivities and civil rights. McQuillan (2019) also draws attention to the ways that AI algorithms shape our subjectivities, dispositions and reliance upon AI technology. Brown (2023) similarly identifies the ways in which AI creates thoughtlessness and structural injustice. Thoughtlessness, as a form of structural injustice, is generated through our increasing dependence on AI. Similarly, Teo (2024) expresses concern that the focus on bias is a form of tunnel vision that ignores broader challenges to protection of individuals from harm and violation of their human rights. Contanza-Chock (2023) raises the question of 'design justice', as a critique of fairness or the elimination of bias. For them, AI violates justice, autonomy and sovereignty and reproduces a colonial epistemology and ontology.

Gender Binarism and Trans and Nonbinary Exclusion in AI

Much of the literature on gender bias does not address the ways in which AI reproduces the gender binary and the consequences of the discriminatory algorithms for trans and nonbinary people (Ricaurte 2022). AI is designed to identify gender from facial data and other categories (Hicks 2019). Constanza-Chock (2020), a nonbinary, trans femme queer person, describes their experiences of humiliation when an airport scanner is unable to clearly discern them as being either male or female. The assumption of everyone having a cisgender identity is built into the algorithms of the scanner.

When the algorithms in machine learning reduce gender to binary gender categories, they render transgender, nonbinary, agender, and intersex people invisible. The assumptions that race, sexuality, and gender are natural biological categories that are fixed and immutable lead to allocating people into categories without their consent. AI is, in effect, actively creating and constructing race, sexuality and gender

categories. While it may not be the intent of machine learning, essentialised biological categories of gender, sexuality and race, which are assumed to be natural and fixed, are being used in AI to justify discrimination and oppression (Crawford 2021).

Dai (2024) demonstrates how AI reproduces ontological sexual difference, gender binarism and patriarchal masculinity within algorithmic binary coding. Within binary coding, gender and sex are seen as immutable, fixed and inevitably binary (Hicks 2019; Katyal & Jung 2021).

Binary gender is reified and presented as an unchangeable category. While historically many feminist scholars differentiated gender from sex, arguing that while gender is socially constructed, and changeable, sex is natural and immutable, this distinction has been increasingly challenged by many feminists since the 1990s (Wittig 1990; Delphy 1993; Fausto-Sterling 2000; Disch 2015; Ruti 2015; Clune-Taylor 2020; Nicholas 2021). Sex, as well as gender, is socially constructed and like gender, is more than binary (D'Ignazio & Klein 2020; Albert & Delano 2021).

The experience of our bodies as sexed is shaped by cultural meanings, social experiences and material arrangements (Clune-Taylor 2020). Gender categories, which are socially constructed, confer gender status on people deemed to have certain biological features (Dembroff 2017). Biology then can be understood as a form of ideology that constructs sex categories such as maleness and femaleness. Biologists and medical scientists project cultural and social meanings onto what are seen as 'objective facts'. It is thus no longer tenable to regard sex as only biological and as an unchangeable base upon which gender is derived (Garlick 2003). Society and the human body are so interconnected that we cannot separate them (Fausto-Sterling 2020). Culture and the social are woven into the biological structure of individual bodies from the beginning (Ruti 2015).

Some authors suggest creating systems that either recognise a third option of sex/gender-neutral alongside male and female, or alternatively eliminate sex/gender as a variable completely. However, various problems have been identified with both options (Albert & Delano 2021). The question is whether AI is capable of operating beyond the male and female binary (Katyal & Jung 2021), while hierarchical gender binary continues to exist in the wider society.

AI, Epistemic Injustice and Gendered Violence

We cannot do justice to the gendered implications of AI without discussing gendered violence. AI has intensified men's harassment and violation of women and nonbinary people online (Aavik et al. 2024). This violence takes many forms, from sexualised deepfake images, non-consensual porn, cyber-stalking, image-based abuse, gender-based trolling, online harassment, and use of AI by perpetrators to control partners (Bailey et al. 2021). In addition, AI is used in predictive policing (Berk 2021), administrative and epistemic violence (Miragoli 2024) and in drone warfare and lethal autonomous weapons (Frantzman 2021).

Deepfake involves an image or video that is created by AI to look like real people saying or doing things that did not happen. Increasingly, deepfakes have enabled image-based abuse, where images of women are superimposed onto nude

women or porn stars engaged in sexual activities (Westerlund 2019). This form of deepfake is a form of image-based sexual abuse and digital violence, and it is a product of generative AI (De Vido 2024). AI-informed deepfake technology has allowed perpetrators new opportunities for sexual exploitation and abuse (Aavik et al. 2024). Bailey et al. (2021) refer to the perpetration of violence, abuse, and harassment online as 'technology-facilitated violence and abuse'. Hall et al. (2024) use the language of 'digital gender-sexual violations' to encompass the multidimensional nature of this violence. Deepfake pornography is one of the many forms of digital violation where women are affected disproportionately (Yangin 2025). In a 2019 study, 99% of deepfake images were of women (Gestoso 2019). One of the biggest challenges in addressing this violence and abuse is the lack of physical contact between the perpetrator and the victim in most cases (Dunn 2021).

A study carried out by Thomas and Balint (2022) demonstrated the ways in which algorithms in YouTube were used to lure boys and young men into the Manosphere, an online collectivity of misogynist men's groups. Ten avatar accounts of young men were algorithmically recommended misogynist and Manosphere content that promoted negative views of feminists and women's rights. Algorithmic structures of websites promote content supposedly to align with users' interests, which in the case of young men, exposes them to extreme viewpoints (Thompson 2023). Currently, in Australia, there is no regulatory framework governing digital platforms to address misogynistic content in algorithms. Beyond the anti-feminism and misogyny of the Manosphere itself (Copland 2015), Nicholas et al. (2024) demonstrate the links between Manosphere content and men's perpetration of domestic violence.

There is also significant ethical concern about the use of AI in law enforcement, and in particular the use of predictive algorithms in policing (Berk 2021). The use of AI in so-called 'pre-emptive policing' perpetuates patriarchal approaches to men's violence against women (Wilcox 2023; Hampton 2023; McInerney 2023) and carceral approaches to violence prevention (McQuillan 2022). The algorithmic models in pre-emptive or predictive policing entrench harmful stereotypes and myths about gendered violence by reinforcing the view that sexual violence is natural and inevitable, and that most sexual violence is committed by strangers in public places. The consequences of this approach are to bolster the model of the state as patriarchal protector, while reinforcing carceral systems of gendered and racialised state violence (McInerney 2023).

Epistemic injustice, an injustice arising from the exclusion and silencing of people's meanings and contributions to knowledge (Fricker 2007), is perpetuated by AI. A number of writers demonstrate epistemic injustices against women and sexual and gender minorities arising from generative AI (Kay et al. 2024; Barry & Stephenson 2025). Kay et al. (2025) use the language of 'algorithmic epistemic injustice' to name the gendered harms of AI. McQuillan (2022) argues that AI perpetuates administrative and epistemic violence, which constitute a form of hermeneutic injustice. He identifies links between the rise of far-right neo-reactionaries and AI data-driven eugenics, which express commitments to innate gender hierarchies. This hermeneutical injustice against women and sexual and gender minorities

perpetuates unequal power dynamics and social prejudices (Barry & Stephenson 2025). Similarly, Teo (2024) uses the language of 'slow violence', a term previously used to describe the incremental ways in which environmental destruction can impact on people (Nixon 2011), to describe the violation of human rights and the gradual harms of AI. Epistemic injustice is thus a form of violence in the ways that it marginalises, exploits, and eliminates belief systems of those it is othering. It is also intertwined with, and facilitates, other forms of gendered violence (Perez & Ramos-Mesa 2025).

From 'Ethics Washing' to an Ethic of Care in AI

Aavik et al. (2024) argue that it is possible to apply feminist principles to AI to develop it in ways that are more aligned with care and social justice. McQuillan (2019) sees potential in feminist technology studies that critique notions of objectivity and neutrality and the promise of standpoint theory that draws attention to political and social privileges associated with particular positions. Ricaurte (2022) argues that a feminist and decolonial imagination is required to develop appropriate ethical guidelines for AI. However, others ask if AI can be feminist, queer, or decolonial (Brown et al. 2023).

Ricaurte (2022) notes that some ethical frameworks constitute a form of 'ethics washing', where they apply narrow technical fixes to problems that are socio-political. Crawford (2021) argues that too often, ethical frameworks are self-regulating and that little attention is given to how principles and ethics should be implemented. More attention should be given to power rather than ethics, and to ensuring that the interests of communities affected by AI should be mobilised to stop its rapid expansion.

One of the negative effects of AI is to increase the moral distance between decision-makers and those impacted by the decision. The algorithms of AI increase anonymity, enable the greater likelihood of unethical behaviour, and come to constitute a form of ethical violence (Villegas-Galavis & Martin 2024). Because there is a lack of transparency in AI decision-making, it obscures the purposes and intentions behind the algorithms. Many of these purposes are unethical and involve manipulation of elections, creating more lethal weapons and increasing global warming and environmental destruction (Cohn 2020).

Various writers propose an ethics of care to address the concerns about AI (Villegas-Galaviz & Martin 2024; Villegas-Galaviz & Fernandez-Fernandez 2022; McQuillan 2022; Cohn 2020; Tailandier 2023). An ethics of care provides a framework to analyse the moral implications of AI. As an alternative, it proposes the importance of interdependence and context to respond to the vulnerability of those people who are impacted by the decision-making process (Villegas-Galaviz & Martin 2024). For McQuillan (2022) also, care is the opposite of the algorithmic detachment of AI because it is focused on dependency and situated vulnerability. However, it is unclear from the authors advocating an ethic of care in AI who will determine the ethics of care and how it can be implemented.

Conclusion: Towards an Alternative Governance of AI

Can AI be democratised? Can there be an AI for the people that is focused on justice rather than discrimination and violence? One proposed strategy is to advocate for greater gender equality and diversity in AI governance (Schopmans & Cupac 2021). The question is whether gender equity and diversity within technology companies, as argued by UN Women (2024) and others (D'Ignazio & Klein 2020; Khan 2024), will effectively address the issues. Wajcman and Young (2023) argue that getting more women and gender-diverse people into science and technology will not address the problem of the embeddedness of the culture of masculinity in the technology itself. Furthermore, greater inclusion and participation by women and gender-diverse others often do not address the power dynamics surrounding technology design and becomes a form of 'participation washing' (Lindgren 2024).

How can people shape the outcomes of algorithmic computation for their own benefit? Bonnie and Trere (2024) suggest that people can use the algorithms to resist the power and interests of those who designed them. Developing what Bonnie and Trere call, 'algorithmic agency', people can decide to resist and break the alliance with data colonialism. They can create online chat groups to share information and educate themselves about how the algorithms work. They can also provide mutual aid and support and organise collective action to subvert and repurpose the algorithms. This is a politics of refusal, opposing the technological inevitability of the current use of AI (Crawford 2021). These forms of everyday resistance, however, are not the solution. While resistance and union organising to address sexism, gender binarism, and racism in the technology labour force is important, there is a deeper problem of the gendered property relations underpinning the technology industry.

How can we find ways of establishing more common ownership? McQuillan (2022) brings the ethic of care into connection with a politics of change that draws upon mutual aid and solidarity to advocate for the practice of commoning. For McQuillan (2021), the way forward is to establish people's councils, a form of federated democratic assembly based on mutual aid and solidarity, as a framework for an alternative approach to AI.

People's councils take the form of what Verdegem (2022) calls the 'digital commons', the communal ownership and distribution of technology and information resources. The digital commons provides an alternative structure for the ownership and governance of AI. It would enable the access and use of data for the common good as opposed to its use as a commodity to serve the interests of patriarchy and big tech. A commons-based practice that is informed by mutual aid and solidarity aims to appropriate resources away from the state and big tech corporations into community ownership. Examples include the free and open software movement (Birkinbine 2018). This would enable shared resources that people would contribute to and use as a common good. However, people's councils may not be enough. To address the masculinism, gender binarism, and whiteness in AI, and to limit vested interests, commoning practices, such as people's councils, will need to draw upon feminist (Clement et al. 2019, queer (Millner-Larsen & Butt 2018) and decolonial (Obeng-Odoom 2021) approaches to the commons to ensure that

different groups and representatives of women, queer folk, and different ethnicities are represented.

To depatriarchalise, queer, and decolonise AI, we will need to challenge the ideology of dataism and its association with masculinism, change the relations of power within the techno-industry by moving the management of AI outside of market relations and state control, and create alternative epistemologies and social imaginaries that foster situated knowledge, social justice, and an ethic of care in the virtual world. Only then will we see the potential for AI to be a force of liberation rather than a technology that perpetuates patriarchy, the gender binary, and colonialism.

Acknowledgements

I would like to thank Jeff Hearn, Russell Shuttleworth, and three anonymous reviewers for their constructive comments on an earlier version of this chapter.

Notes

1 It is also important to acknowledge the contributions of First Nations voices in AI. See, for example, Silano (2024).
2 The focus of this chapter is on the development of AI in the global North. I acknowledge that China also has ambition to be a global AI superpower (Cheng & Zeng 2023). However, as these developments are taking place within China's state-centric, sovereignty-oriented approach, the strategies for addressing masculinism and gender binarism would take different forms and I cannot do justice to how these issues unfold in state-controlled environments.

References

Aavik, K., Collinson, D. L., Hall, M., Hearn, J., & Thym, A. (2024). The impact of men's domination of AI and deepfake technology: Violations of women, girls, and democracy in a digital age. *Routledge Blog.* https://oru.divaportal.org/smash/get/diva2:1901409/FULLTEXT01.pdf

Abbate, J. (2012). *Recoding gender: Women's changing participation in computing*. MIT Press.

Albert, K., & Delano, M. (2021). 'This whole thing smacks of gender': Algorithmic exclusion in the bioimpedance-based body composition analysis. In *FAccT Proceedings of the ACM Conference on Fairness, Accountability and Transparency*, pp. 342–352. https://doi.org/10.1145/3442188.3445898.

Bailey, J., Henry, N., & Flynn, A. (2021). Technology-facilitated violence and abuse: International perspectives and experiences. In J. Bailey, N. Henry & A. Flynn (Eds.), *The emerald international handbook of technology-facilitated violence and abuse* (pp. 1–17). Emerald Publishing.

Barry, I., & Stephenson, E. (2025). The gendered epistemic injustices of generative AI. *Australian Feminist Studies 40*(123), 1–21. https://doi.org/10.1080/08164649.2025.2480927.

Basset, C.e (2023). The cruel optimism of technological dreams. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 258–173). Oxford University Press.

Berk, R. (2021). Artificial intelligence, predictive policing and risk assessment for law enforcement. *Annual Review of Criminology, 4,* 209–237. https://doi.org/10.1146/annurev-criminol-051520-012342

Birkinbine, B. (2018). Commons praxis: Towards a critical political economy of the digital commons. *Triple C: Communication, Capitalism & Critique, 16*(1), 290–305. https://doi.org/10.31269/TRIPLEC.V16I1.929.

Blas, Z., Jue, M., & Rhee, J. (2025). Introduction: Chart of transitions. In Z. Blas, M. Jue & J. Rhee (Eds), *Informatics of domination* (pp. 1–31). Duke University Press.

Bonni, T., & Trere, E. (2024). *Algorithms of resistance: The everyday fight against platform power.* MIT Press.

Brittan, A. (1989). *Masculinity and power.* Willey-Blackwell.

Brown, J., Cave, S., Drage, E., & McInerney, K. (2023). Introduction: Feminist AI. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. xii–xxi). Oxford University Press.

Browne, J. (2023). AI and structural injustice: A feminist perspective. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 328–346). Oxford University Press.

Butcher, M., Cohen, E., Kunkle, C., & Totzkay, D. (2022). Geek girl today, scientist tomorrow? Inclusive experiences and efficacy mediating the link between women's engagement in popular geek culture and STEM career interest. *International Journal of Science Education Part B, 13*(3), 276–291. https://doi.org/10.1080/21548455.2023.2172624.

Cheng, J., & Zeng, J. (2023). Shaping AI's future? China in AI's global governance. *Journal of Contemporary China, 32*(143), 798–810. https://doi.org/10.1080/10670564.2022.2107391.

Clement, F., Harcourt, W., & Joshi, D. (2019). Feminist political ecologies of the commons and commoning. *International Journal of the Commons*, *1*, 1–15. https://doi.org/10.18352/ijc.972.

Clune-Taylor, C. (2020). Is sex socially constructed? In S. Crasnow & K. Intemann (Eds), *Routledge handbook of feminist philosophy of science* (pp. 293–315). Routledge.

Cohn, J. (2020). In a different code: Artificial intelligence and the ethics of care. *International Review of Information Ethics*, *28*, 1–7.

Copland, S. (2025). *The male complaint: The manosphere and misogyny online*. Polity.

Costa, P., & Ribas, L. (2019). AI becomes her: Discussing gender and artificial intelligence. *Technoetic Arts: A Journal of Speculative Research, 17*(1&2), 171–193. https://doi.org/10.1386/tear_00014_1.

Contanza-Chock, S. (2023). Design practices: 'Nothing about us, without us'. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 370–388). Oxford University Press.

Couldry, N., & Mejias, U. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media, 20*(4), 336–349. https://doi.org/10.1177/1527476418796632.

Crawford, K. (2021). *Atlas of AI: Power, politics and planetary costs of artificial intelligence.* Yale University Press.

D'Ignazio, C., & Klein, L. (2020). *Data feminism*. MIT Press.

D'Ignazio, C., & Klein, L. (2023). The false binary of reason and emotion in data visualisation. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 193–207). Oxford University Press.

Dai, Yongde (2024). Deconstructing Ex Machina (2014): A feminist-psychoanalytic exploration of female artificial intelligences. *Frontiers in Communication, 9*. https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2024.1349874/full.

Delphy, C. (1993). Rethinking sex and gender. *Women's Studies International Forum, 16*(1), 1–9.
Dembroff, R. (2017). *Categories we (aim to) live by* [PhD Thesis]. Faculty of Princeton University.
De Vido, S. (2024). Deep fake as AI-generated violence against women. https://www.unive.it/pag/fileadmin/user_upload/dipartimenti/DSLCC/documenti/DEP/finestra/Finestra_19.pdf
Disch, L. (2015). Christine Delphy's constructivist materialism: An overlooked 'French feminism'. *South Atlantic Quarterly, 114*(4), 827–849. https://doi.org/10.1215/00382876–3157155.
Drage, E., & Frabetti, F. (2023). AI that matters: A feminist approach to the study of intelligent machines. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 274–289). Oxford University Press.
Dunn, S. (2021). Is it *actually* violence? Framing technology-facilitated abuse as violence. In J. Bailey, N. Henry & A. Flynn (Eds.), *The emerald international handbook of technology-facilitated violence and abuse* (pp. 25–45). Emerald Publishing.
Dwiputri, A. (n.d.). *Gender bias in artificial intelligence: The role of developers in reproducing inequality*. Asia Pacific Foundation of Canada.
Fausto-Sterling, A. (2020). *Sexing the body: Gender politics and the construction of sexuality.* Hachette.
Frantzman, S. (2021). *Drone wars: Pioneers, killing machines, artificial intelligence and the battle for the future.* Post Hill Press.
Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford.
Garlick, S. (2003). What is a man? Heterosexuality and the technology of masculinity. *Men and Masculinities, 6*(2), 156–172. https://doi.org/10.1177/1097184X03255851.
Gestoso, P. (2019). Techno-patriarchy: How AI is misogyny's new clothes. *Galvanizing Diversity and Inclusion through Technology.* https://patriciagestoso.com/2024/01/08/techno-patriarchy-how-ai-is-misogynys-new-clothes/
Hall, M., Lewis, R., & Hearn, J. (2014). The bounded limitlessness of digital gender-sexual violations: The implications for women and gender-sexual relations. *Violence Against Women*, 1–22. https://doi.org/10.1177/10778012241292293.
Hampton, L. (2023). Techno-racial capitalism: A decolonial black feminist Marxist perspective. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 119–136). Oxford University Press.
Hicks, M. (2019). Hacking the Cis-tem, *IEEE Annals of the History of Computing, 41* (1), pp. 20-33, doi: 10.1109/MAHC.2019.2897667.
Hoffman, A. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse, *Information, Communication & Society, 22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912.
Katyal, S., & Jung, J. (2021). The gender panopticon: AI, gender and design justice. *UCLA Law Review, 2021*, 692–787.
Kay, J., Kasirzadeh, A., & Mohammed, S. (2025). Epistemic injustice in generative AI. *Proceedings of the AAA/ACM Conference on AI, Ethics and Society*, https://dl.acm.org/doi/10.5555/3716662.3716722, pp. 684–697.
Keeling, D. (2012). History of (future) progress: Hyper-masculine transhumanist virtuality. *Critical Studies in Media Communication, 29*(2), 132–148. https://doi.org/10.1080/15295036.2012.666803.
Khan, R. (2024). Patriarchal AI: How ChatGPT can harm a woman's career. *LSE Research Online* https://eprints.lse.ac.uk/123478/1/medialse_2024-4-26--patriarchal-ai-how-chatgpt-can-harm-a-womans-career.pdf.

Lindgren, S. (2024). *Critical theory of AI*. Polity

Little, B., & Winch, A. (2021a). *The new patriarchs of digital capitalism*. Taylor and Francis.

Little, B., & Winch, A. (2021b). Patriarchy in the digital conjuncture: An analysis of Google's James Damore. *New Formations: A Journal of Culture/Theory/Politics, 102*, 44–63. https://doi.org/10.3898/NEWF:102.04.2020.

Lockhart, E. (2015). *Nerd/geek masculinity: Technology, rationality and gender in nerd culture's countermasculine hegemony* [Doctor of Philosophy Thesis]. Texas A&M University.

McInerney, K. (2023). Coding 'carnal knowledge' into carceral systems: A feminist abolitionist approach to predictive policing. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 101–118). Oxford University Press.

McQuillan, D. (2019). Non-fascist AI. In M. Hlavajova & W. Maas (Eds.), *Propositions for non-fascist living: Tentative and urgent* (pp. 113–124). BAK basis voor actuele kunst and MIT Press.

McQuillan, D. (2021). Post-humanism, mutual aid. In P. Verdegen (Ed.), *AI for everyone? Critical perspectives* (pp. 67–84). University of Westminster Press.

McQuillan, D. (2022). *Resisting AI: An anti-fascist approach to artificial intelligence*. Bristol University Press.

Mendz, G., & Cook, M. (2021). Transhumanist genetic enhancement: Creation of a 'new man' through technical innovation. *New Bioethics*, *27*(3), 105–126. https://doi.org/10.1080/20502877.2021.1917228.

Millner-Larsen, N., & Butt, G. (2018). Introduction: The queer commons: *GLQ: A Journal of Lesbian and Gay Studies, 24*(4), 399–419. https://doi.org/10.1215/10642684-6957744.

Miragoli, M. (2024). Conformism, ignorance and injustice: AI as a tool of epistemic oppression. *Episteme, 22*(2), 522–540. https://doi.org/10.1017/epi.2024.11.

Mohla, S., Bagh, B., & Guha, A. (2024). *Thinking beyond bias: Analysing impacts and implications of AI on gendered labour*. arXiv:2406.16207, https://doi.org/10.48550/arXiv.2406.16207.

Mollema, W. (2024). Decolonial AI as disenclosure. *Journal of Social Issues, 12*, 574–603. https://doi.org/10.4236/jss.2024.122032.

Nicholas, L. (2021). Remembering Simone de Beauvior's 'Ethics of ambiguity', to challenge contemporary divides: Feminism beyond both sex and gender. *Feminist Theory, 22*(2), 226–247. https://doi.org/10.1177/1464700120988641.

Nicholas, L., Clark, S., Agius, C., & Cook, K. (2024). Antifeminist, manosphere and right-wing extremist sentiment among men who use domestic and family violence: Masculinism, misinformation and the justificatory logics of violence. *NORMA: International Journal for Masculinity Studies*, 1–20. https://doi.org/10.1080/18902138.2024.2430513.

Nixon, R. (2011). *Slow violence and the environmentalism of the poor*. Harvard University Press.

Noble, S. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.

Obeng-Odoom, F. (2021). *The commons in an age of uncertainty: Decolonizing, nature, economy and society.* University of Toronto Press.

Perez, M., & Ramos-Mesa, A. (2025). Epistemic violence: An analytical tool for theorising the interconnections of violences. In K. Seymour, B. Pease, S. Strid & J. Hearn (Eds.), *Interconnecting the violences of men: Continuities and intersections in research, policy and activism* (pp. 248–264). Routledge.

Ricaurte, P. (2022). Artificial intelligence and the feminist decolonial imagination. https://botpopuli.net/artificial-intelligence-and-the-feminist-decolonial-imagination/

Richards, P. (2020). Big data and the resurgence of the far right within the United States of America. In S. Ashe, J. Busher, G. Macklin & A. Winter (Eds.), *Researching the far right: Theory, method and practice*. (pp. 164–177). Routledge.

Ruti, M. (2015). *The age of scientific sexism*. Routledge.

Salter, M. (2018). From geek masculinity to gamergate: The technological rationality of online abuse. *Crime, Media, Culture, 14*(2), 247–264. doi.org/10.1177/1741659017690893.

Schopmans, H., & Cupac, J. (2021). Engines of patriarchy: Ethical artificial intelligence in times of illiberal backlash politics. *Ethics & International Affairs, 35*(Fall), 329–342. https://doi.org/10.1017/S0892679421000356.

Silano, J. (2024). Towards abundant intelligences: Considerations for Indigenous perspectives in adopting artificial intelligence technology. *Healthcare Management Forum, 37*(5), 329–333. https://doi.org/10.1177/08404704241257144.

Tailandier, A. (2023). AI in a different voice. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 32–46). Oxford University Press.

Teo, S. (2024). Artificial intelligence and its 'slow violence' to human rights. *AI and Ethics*. https://www.researchgate.net/publication/383235625_Artificial_intelligence_and_its_'slow_violence'_to_human_rights.

Thomas, E., & Balint (2022). *Algorithms as a weapon against women: How YouTube lures boys and young men into the 'manosphere'*. Institute for Strategic Dialogue.

Thompson, C. (2023). *Digital masculinity: An analysis of how masculine values are manifested in online spaces. Honors Project*. Seattle Pacific University.

UN Women (2024). Artificial intelligence and gender equity. https://www.unwomen.org/en/articles/explainer/artificial-intelligence-and-gender-equality#:~:text=A%20study%20by%20the%20Berkeley,both%20gender%20and%20racial%.

Varon, J., & Pena, P. (2021). Artificial intelligence and consent: A feminist anti-colonial critique. *Internet Policy Review, 10*(4), 1–25. https://doi.org/10.14763/2021.4.1602.

Verdegem, P. (2022). Dismantling AI capitalism: The commons as an alternative to the power centration of Big Tech. *AI & Society, 39*, 727–737. https://doi.org/10.1007/s00146-022-01437-8.

Villegas-Galaviz, C., & Fernandez-Fernandez, J-L. (2022). Care ethics in the era of artificial intelligence. In G. Faldetti, E. Mollona & M. Pellegrini (Eds.), *Philosophy and business ethics: Organizations, CSR and moral practice* (pp. 255–274). Palgrave Macmillan.

Villegas-Galaviz, C., & Martin, K. (2024). Moral distance, AI and the ethics of care. *AI & Society, 39*, 1695–1706. https://doi.org/10.1007/s00146-023-01642-z.

Wajcman, J. (1991). *Feminism confronts technology*. The Pennsylvania University Press.

Wajcman, J., & Young E. (2023). Feminism confronts AI: The gender relations of digitalisation. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 47–64). Oxford University Press.

Walther, A., Logoz, F., & Eggenberger, L. (2024). The gendered nature of AI: Men and masculinities through the lens of ChatGPT and GPT4. *Computers in Human Behavior: Artificial Humans*, *2*(2), 1–7.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52. https://doi.org/10.22215/timreview/1282.

Wilcox, L. (2023). No humans in the loop: Killer robots, race and AI. In J. Browne, S. Cave, E. Drage & K. McInerney (Eds.), *Feminist AI: Critical perspectives on data, algorithms and intelligent machines* (pp. 83–100). Oxford University Press.

Wilshire, C. (2014). *Time to reboot: Feminism in the algorithm age*. Monash University Publishing.

Wittig, M. (1990). *The straight mind and other essays*. Harvester Wheatsheaf.

Yangin, L. (2025). Deepfakes and the manosphere: A dangerous combination. https://she.org.sg/news/deepfakes-the-manosphere-a-dangerous-mix

10 The Helping Relationship in New Online Relational Spaces

AI Chatbots and the Reconfiguration of Trust in Social Work

Luca Pavani and Mara Sanfelici

Introduction: Digital Transformation Reconfiguring Social Work Practices

The helping relationship in social work has been traditionally understood as an interpersonal, situated, and ethically grounded encounter between a service user and a professional. It involves being grounded in one's self, while being responsive and attuned to the other, being fully in the moment, on a multitude of levels: physically, emotionally, cognitively, relationally, and spiritually. This relationship is not just a means to optimise and collaboratively perform tasks (information exchange, assessment, service delivery), but it is in itself a way of caring.

In the last decades, this relational space has been increasingly populated by new technologies that allow communication from a distance involving different senses. New ways of being connected have been made possible by the use of phones, chat, or videocalls, that help to overcome traditional barriers in accessing services, such as geographical distances, and stigma associated with seeking help. In online interaction, however, physical distance limits nonverbal communication cues (Oshni Alvandi, 2019) and in such settings both social workers and service users need to find new ways to express presence and build trust (Bos et al., 2002). New processes of connection and disconnection have pros and cons associated with the use of such technologies, that – for the workers – require not only the learning of new communication techniques, but also a constant process of reflexivity regarding the methodological and ethical implications of such transformations (Sanfelici and Nothdurfter, 2026).

More recently, new AI chatbots have been designed to interact directly with service users, without the presence of a professional. Available 24/7, able to serve unlimited users simultaneously, and accessible from every personal computer or smartphone, they have been described as new means that promise to overcome physical, economical, and cultural barriers in accessing social care (Victor and Goldkind, 2025). These systems can provide information, guidance, emotional support, or preliminary assistance, sometimes in place of, and sometimes alongside, human professionals. This shift marks a significant transformation: the helping relationship is no longer exclusively human-to-human, but increasingly human-to-machine.

DOI: 10.4324/9781003627920-10
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

This chapter focuses on the implications of this ongoing transformation, asking how the encounter between service users and AI chatbots differs from, overlaps with, reconfigures or replaces the traditional professional–user relationship, and what may be the consequences for the field of social work. Rather than discussing the influence of AI in terms of (in)efficiency or (in)accuracy, the analysis concentrates on the relational qualities of help. Are machines able to deliver equal or superior outcomes than human professionals, as it is suggested by a strand of literature (Victor Goldkind, 2025). Is there, instead, a unique value that social workers provide in a helping relationship or, on the other hand, a unique contribution of such new technologies? To answer these questions, we will refer to the available theoretical and empirical literature in social work and related disciplines, that help to reflect on the main differences between a human-to-human and a human–machine interaction. In particular, we are going to focus on how people and technologies interact in the process of trust-building, that highly influences the quality of a helping relationship. How and why do service users trust professionals and when and for which reasons do they instead prefer to trust and be helped by AI chatbots?

The assumed theoretical lens that guides this work is a relational and sociomaterial perspective that considers technologies not just as tools with an impact, but looks at how human and non-human elements interact within fluid assemblages, that at the same time create and reconfigure structures (Orlikowsky, 2000). In this perspective, mediatisation is not defined as "a media-centric postulate about the media's direct 'causal influence' on every aspect of society" (Hepp et al., 2015: 316), but as a complex network of interrelations between various cultural, technological, political, and economic aspects and their transformative potentials, within power relations.

Against this backdrop, the following paragraphs examine how trust is constructed within professional helping relationships and how its meaning and conditions are reconfigured in interactions with AI chatbots, in order to critically assess what is gained, what is lost, and what is transformed when help is increasingly mediated by socio-technical systems rather than by human relationships.

Trust in the Helping Relationship

In a helping relationship, building a connection based on trust is the way to allow a reciprocal understanding of each other and the situation – as the basis of the assessment process; moreover, trust is a necessary condition to foster participation and exchange, therefore to co-build possible paths and imagine solutions to tackle conditions of socially constructed forms of disadvantages and discriminations. In this process, trust is progressively built and reassessed over time.

There is no unique definition of trust across disciplines, and – despite its centrality – the concept's meaning is often taken for granted in the social work literature. Behnia (2008) specifically reflected on the process of trust-building, starting from its definition as one's willingness to be vulnerable to others based on one's positive expectations of the other's intention and competence (Blomqvist and Stahle, 2004; Levin et al., 2006), This author focuses on an essential factor related to trust:

the exposure to vulnerability, a core topic in social work. Service users are expected to share very personal information about their lives and their feelings, while the social workers' effort to understand the clients, trying to 'see' and 'feel' from their perspectives, and being there for the other in an authentic way, is based on an attitude of openness that implies personal exposure on the professional's side also.

ln trusting others, first of all, one expects that one's vulnerability will not be exploited for personal interest or reason of power (Pellegrino, 1991), but that, on the contrary, the other will care for them or at least will not do any harm (Northhouse and Northhouse, 1985; Levin et al., 2006).

In this definition, trust also involves a positive expectation about the other's ability and competence to accomplish something that has motivated the encounter (Levin et al., 2006; Blomqvist and Stahle, 2004).

Considering that trust is based on reciprocal expectations and perceptions, more or less confirmed during the encounters, it cannot be considered as a fixed attribute of an interaction, nor can it be reduced to individual characteristics or rational calculations.

Several theories have been proposed to interpret the process of trust-building. Each of them captures partial aspects that should be read within a relational and systemic perspective. Dispositional approaches have framed trust as a stable personality trait; people identified as low or high trusters influence differently the outcome of the relationship in terms of trust. These approaches also highlight an internalised more positive orientation towards people who share similar characteristics, based on common gender, class, ethnicity, or professional backgrounds. A different model of explanation locates trust in the perceived features of the people considered trustworthy, such as credibility, competence, honesty, warmth, openness, acceptability of opinion, and reliability.

All these perspectives are useful to account for some of the involved dynamics, even if they fail to explain how trust is constructed, challenged, and transformed over time in real interactions (Behnia, 2008). Contemporary models of explanation converge on the view that trust should be understood as a situated social process, temporally distributed and constructed through relationships (Khodyakov, 2007; Eyal et al., 2024).

In fact, trust develops through ongoing social interaction, shaped by how actors interpret one another's actions, intentions, and meanings within specific relational and institutional contexts. Drawing on symbolic interactionism, Behnia (2008) emphasises that individuals act toward others based on the meanings those others hold for them, and that these meanings are produced and revised through social interaction, trust, therefore, is grounded in the interpretation of interactional cues and processes of co-construction of meaning.

In client-professional relationships, the first encounter is the space where trust can be built to initiate a relationship. Clients need to have confidence in the intentions and competencies of the social workers, in relation to the problem they are trying to make sense of. This confidence is built within a continuously negotiated interaction, in which several variables are at stake, including the asymmetrical power in the client-professional relationship. Trust is both a condition for the engagement and a

relational achievement filled with tensions: the actors involved are required to engage despite uncertainty and, sometimes, fear; at the same time, it implies openness that exposes them to vulnerability, and a certain degree of certainty.

In fact, trust-building involves at least three processes. Firstly, trust develops through concrete practices, understood as actions and feedback that enable the actors to assess one another's reliability over time, thereby rendering the other's behaviour interpretable and, consequently, predictable (Eyal et al., 2024). In the helping process, trust can develop when service users interpret professionals' actions as consistent, responsive, and caring, and when these interpretations are confirmed over time through interactions. Through repeated encounters, service users test the professional's interpretations, observe patterns of behaviour, and assess whether the relationship allows for the expression of vulnerability, without undue risks. Trust, in this sense, is reinforced when social workers demonstrate consistent embodied presence, capacity to listen and care for, acknowledge uncertainty and ambiguity, and engage reflexively with the client's perspective, enabling a form of interaction in which trust can deepen over time. Secondly, trust is sustained by shared interpretations, which arise when the parties involved succeed in constructing a sufficiently common definition of the situation, the respective roles, and the purposes of the interaction, even in the absence of full convergence or control over outcomes (Khodyakov, 2007). Thirdly, trust relies on future-oriented expectations that make action feasible under conditions of partial knowledge, as trust allows doubts to be suspended within tolerable limits and socially manageable risks to be undertaken (Luhmann, 1979).

This process of trust-building is not linear or guaranteed. Trust can be fragile and contested, and unfolds within structurally asymmetric relationships; first, clients and professionals occupy differentiated positions of power, vulnerability, and responsibility; moreover, both clients and professionals do not interact in a vacuum but in networks in which power relations unfold, influencing processes of inclusion or exclusion that highly influence the possibility of reciprocal trust.

From the client's perspective, entering a relationship with a social worker involves an act of entrustment, accepting a condition of exposure that entails the sharing of intimate information, and the entry into evaluative and decision-making processes, that can be more or less participatory and that in several cases may affect rights and life trajectories. This act of entrusting therefore takes place in contexts marked by uncertainty and vulnerability. On the other hand, social workers have more power and discretion, given their professional status and the possibility to affect the access to resources and rights (Lipsky, 2010). For an authentic process of trust-building, such power needs to be balanced through practices of listening, clarity regarding the professional mandate, negotiation of aims, and the demonstration of the concrete usefulness of professional action, situating trust-building within dynamic and reversible processes rather than treating it as a given prerequisite (Trevithick, 2003; Ruch et al., 2018). This entails the intentional recognition of the client as a subject endowed with resources, interpretive capacities, and experiential competences that are relevant to understanding the situation and defining the goals of the intervention.

Social workers are also potentially vulnerable: their actions can be questioned, their interpretations challenged, and they may work within institutional settings oriented more toward social control than care, or organised according to criteria that can either enable or constrain relational practice. Eliminating the risk of misunderstanding is impossible, but the possibility to build dialogue, exchange, and clarification in a safe environment is what makes the difference when tensions arise. Service users develop trust not simply because professionals possess expertise or institutional authority, but because they experience interactions as meaningful, and oriented toward shared understanding, and have the possibility to feel valued and recognised in this interaction, even when institutional constraints remain (Behnia, 2008).

This process also unfolds involving the physical expression of bodies. Professional's embodiment and expression of their grounded presence can facilitate a sense of safety through a process of co-regulation, which allows connection. Non-verbal cues (prosodic vocal tone, gesturing, having an open body posture) are used to communicate responsiveness and presence, with the aim of providing a space where the person can feel safe, felt, and understood. Reciprocal trust is also the result of synchronisation of physiological rhythms, bodily movements, and cognitive attunement involved in reciprocal processes of meaning-making.

The next paragraph focuses on how the introduction of artificial intelligence in the interaction between citizens and welfare services can transform such processes, looking at how AI chatbots are creating new relational spaces as new environments in which humans learn, express and exchange knowledge, feelings, and experiences in new ways, with concrete consequences for social work practice.

Trust in AI Chatbots

Since AI chatbots have entered social work practice, they are being experienced as useful and, in some cases, preferable tools of support, a key question nowadays concerns how AI chatbots reconfigure the conditions under which processes of trust-building are shaped and more generally how different forms of help are experienced and made meaningful.

A chatbot is a computer program designed to simulate conversation with humans (Chan and Li, 2023), trained to recognise their intent and the most appropriate response to their request, based on predetermined answers and contextualised analysis of data. Currently, chatbots use different methods to implement such conversation cycles (Chan and Li, 2023). Most existing chatbots are retrieval-based: they typically involve machine learning to train the system to understand various intents and retrieve preset answers from a database.

Trust in AI chatbots introduces a partially distinct set of dynamics, when compared to human-to-human interactions. In the field of technologies and AI studies, trust is generally defined as a condition that enables acceptance and engagement with a machine and, more specifically, it is referred as the willingness to rely on a technological system, while accepting a degree of risk and uncertainty concerning its outputs (Afroogh et al., 2024). Therefore, like in interactions among

humans, vulnerability is considered an important variable involved in this process of trust-building; trust in AI is shaped by factors such as perceived accuracy, reliability, transparency, and explainability, which contribute to making system behaviour more predictable and to reducing users' cognitive uncertainty (Afroogh et al., 2024; Gursoy et al., 2019; Becker and Fischer, 2024).

Nevertheless, this form of trust rests on a functional reciprocity of interaction, whereby the system responds to users' inputs and adapts to them, while remaining devoid of intentionality and moral agency.

Critically reflecting on this main difference between trust in AI and the trust that underpins professional–client relationships helps to understand which are considered meaningful interactions, and what counts as care and support in contemporary welfare systems.

Within social work, discussion on conversational AI systems remains limited and the available empirical literature tends to focus on narrowly defined clinical applications, including smoking cessation interventions (He et al., 2022), self-help tools for depression (Liu et al., 2022), follow-up support after cancer treatment (Greer et al., 2019), and experimental efforts to design chatbots for social work education and professional training (Chan and Li, 2023). Broader theoretical questions concerning trust, relational dynamics in human–chatbot interactions, and the context in which they are shaped remain under-theorised within the field, despite the importance of such topics. These systems have been made to mediate, support, or replace relational functions traditionally exercised by professionals, therefore methodological and ethical discussions on the consequences of such reconfiguration of help are needed.

Through the lens of a relational perspective (outlined above), trust in AI is not viewed as a single dyadic relationship, but as a phenomenon emerging from the interaction of institutional choices, professional practices, and clients lived experiences of human–machine interactions, implying a potentially profound reconfiguration of the conditions under which help is accessed.

AI chatbots' appeal lies in a complex assemblage of practical affordances and relational effects. A recent literature review (Ng and Zhang, 2025) shows how trust in AI chatbots is shaped by user-related, machine-related, interaction-related, socio-cultural, and contextual variables, interacting within socio-material assemblages. The characteristics of the technology in use and their affordances matter: information quality, data security, and possibilities for personalisation emerge as central variables, together with machines' human-like characteristics, such as the ability to formulate empathetic responses and anthropomorphic cues. These features support both cognitive and affective forms of trust by allowing interactions to be experienced as socially and emotionally meaningful. Empirical findings concerning user-related factors are more heterogeneous (Ng and Zhang, 2025). Demographic variables (gender, age) generally show limited explanatory power, while higher levels of technical knowledge or domain expertise may, in some cases, reduce trust due to more critical assessments of system performance. Previous experience with chatbots tends to be associated with higher trust, although familiarity alone does not appear sufficient to foster trust, while perceived risk, particularly

with respect to privacy, is associated with lower levels of trust. Context-related factors refer to conditions surrounding the usage of the machine and more generally social representations of these technologies, that influence trust and the overall users' experience.

Trust is associated not only with users' or machines' characteristics, but also with what emerges as socio-technical assemblages, that in some ways seem to attenuate some of the most threatening dimensions of the helping relationship, at least in its initial phase, as well as to respond to long-standing frictions within welfare systems, particularly those shaped by austerity, bureaucratisation, and stigma. As Khawaja and Bélisle-Pipon (2023) argue, chatbots reduce various barriers to accessing social services, by eliminating the need for physical mobility, scheduled appointments, and direct exposure to institutional environments. These settings are not neutral backdrops, but historically and symbolically charged spaces, sometimes associated with surveillance, moral judgement, and asymmetrical power relations. By allowing users to seek support from private or familiar settings, chatbots seem to re-spatialise welfare encounters, "protecting" users from the physical and symbolic "architectures" that have traditionally shaped help-seeking. This spatial displacement is accompanied by temporal flexibility: the availability of chatbots at any time helps align the emergency of needs and the possible responses, overcoming issues related to the actual availability of professionals, and the experience of "waiting for" uncertain answers from social services. For many users, particularly those experiencing crisis, this continuous accessibility is not merely convenient but constitutive of the experience of being "seen" and recognised as important. Research on mediated interactions and social presence has consistently shown that immediacy and responsiveness are central to perceived recognition; in this sense, constant availability functions as a relational cue, signalling attentiveness and care even in the absence of a human interlocutor.

These practical gains are inseparable from a third dimension: the reconfiguration of control within the interaction. Chatbots allow users to engage at their own time and their own pace, to pause, repeat, reformulate, or withdraw without social consequences. This capacity to regulate the depth of interaction is repeatedly identified in human-computer interaction research as a key factor in perceived safety and trust. When users experience the interaction as controllable and predictable, it becomes subjectively less demanding and less risky, particularly in the early stages of help-seeking when uncertainty and fear of exposure are more acute. Importantly, this sense of control is not only technical but symbolic: it allows users to maintain a degree of agency in contexts where it has often been eroded by institutional procedures. In fact, the capacity to listen to, give time and space to the other is the base of a relational social work practice. However, what not rarely happens is that human-professional encounters in social work are structured by time constraints, role expectations, and institutional scripts that can delimit what is appropriate to talk about. Even in relationally oriented practice, clients often experience pressure to articulate coherent narratives, manage emotions, and conform to implicit norms of deservingness. The perceived absence of judgement further amplifies a sense of trust. The possibility of regulating self-disclosure – deciding what to say, how

much to say, and when to say it – lowers symbolic barriers to engagement and can make initial contact with support services more likely. Moreover, welfare encounters take place in contexts often marked by stigma, moral evaluation, and anticipatory shame, particularly for individuals experiencing multiple forms of social oppression. For such users, interaction with an automated system may feel safer precisely because of the availability of an environment potentially "liberated" from the effect of socially constructed categories shaping discrimination. The chatbot does not appear to scrutinise body language, tone, or affect, nor focus on marks of poverty, race, or diverse physical appearance. While this neutrality is, in practice, a carefully designed effect rather than an inherent property of the system, it nonetheless shapes users' lived experiences.

The appeal of chatbots cannot be fully explained by accessibility, control, and non-judgement alone. Recent studies suggest that natural language-based systems are capable of eliciting strong emotional responses and fostering forms of engagement that subjectively resemble human relational dynamics (Grassi, 2025). Through conversational fluency, continuity across interactions, and the simulation of attentiveness, chatbots can produce a sense of being listened to and recognised. Conversational AI can operate as a tool that produces trust-inducing authenticity (Betlemidze, 2022); responses sound eerily sincere, emotionally attuned, and personally vulnerable, despite being generated algorithmically. Design choices such as handwriting-like typefaces, personalised greetings, and adaptive language may foster a sense of intimacy, inviting users to interpret the interaction as relational rather than purely instrumental. These cues sustain affective engagement and can keep users bonded within what Betlemidze conceptualises as processes of "trans-corporeal surrogacy", where meaningful connections, satisfaction, and comfort are experienced without reciprocal human presence.

In recent theorisation of human–machine communication, the concept of enchantment has been introduced to account for these affective, sensory, and relational dynamics that emerge in encounters between humans and artificial intelligence. Betlemidze (2022) explicitly distances the notion of enchantment from naïve technological fascination or uncritical celebration of innovation. Instead, enchantment is theorised as a relational and processual condition that temporarily suspends habitual modes of perception and interpretation, allowing new forms of connection and meaning to arise, also within mediatised socio-material assemblages. Enchantment as a form of "acute sensory activity" is a pre-reflexive, affective state that does not have to involve rational evaluation or conscious belief, in which individuals experience wonder and surprise within an embodied openness that precedes cognitive judgement, enabling actors to momentarily engage with phenomena outside established interpretive frameworks. Betlemidze extends this understanding by emphasising that enchantment is not confined to human-human interaction, but can arise between humans, machines, and environments. In particular, she shows through examples how the human-like voice of an AI chatbot produces a presence that is simultaneously immediate and mediated, intimate and disembodied, constituting an encounter that precedes visual recognition and categorisation. The voice appears personal, responsive, and emotionally attuned, even

while its machinic nature remains partially opaque, thereby sustaining a space of ambiguity in which enchantment can flourish.

Importantly, enchantment in this framework does not equate to trust, nor should it be understood as a stable or morally grounded relational state. Rather, enchantment functions as a condition of possibility for subsequent relational processes, including trust, intimacy, and attachment. Betlemidze explicitly situates enchantment at an earlier stage of relational development, where affective openness enables vulnerability, neediness, and the perception of authenticity to emerge. Enchantment allows actors to momentarily bracket scepticism and instrumental reasoning, making it possible to engage with the AI as a relational presence rather than as a mere tool. In doing so, enchantment prepares the ground for trust without guaranteeing it. Trust, if it develops, does so later and through additional processes of interaction, interpretation, and reliance.

This concept is useful to understand human-AI relationships, particularly in contexts of care, support, and emotional engagement. Enchantment explains why individuals may initially feel drawn to, comforted by, or emotionally connected with AI systems, even before forming explicit judgements about their reliability or ethical status. The affective intensity of enchantment renders the interaction meaningful at a sensory and experiential level, thereby facilitating forms of engagement that cannot be reduced to cognitive evaluation alone. At the same time, because enchantment is pre-reflexive and affect-driven, it is inherently fragile and reversible. In fact, enchantment is always accompanied by the possibility–and eventual inevitability–of disenchantment, both in human-to-human and human–machine relationships. Disenchantment occurs when instrumental modes of understanding reappear and when AI machines are involved, with the reappearance of a world governed by calculation and predictability models.

Enchantment helps explain how relational engagement with AI systems can begin prior to trust, moral judgement, or institutional legitimation, and why such engagement may feel deeply meaningful despite its fragility. At the same time, the concept resists explanations that limits the focus to forms of illusion or manipulation, instead situating enchantment within broader assemblages of mediation, affect, and power relations, helping to critically analyse how such assemblages emerge, are shaped, and reconfigure a new concept of being and feeling helped.

Firstly, in chatbots–human interactions, trust becomes anchored in the subjective experience of language and emotional resonance, rather than in genuine reciprocity or moral accountability. The system appears to "listen, understand, and know" not because it possesses human understanding, but since it performs these qualities convincingly at the level of interaction. This performance is sufficient, in many cases, to generate for humans a sense of being recognised, particularly when contrasted with prior experiences of rushed, impersonal, or judgement-laden professional encounters. In this sense, chatbots may respond to an unmet relational need within welfare systems: the desire for attention, continuity, and acknowledgement that is sometimes experienced as lacking within assessment or decision-making processes in social services.

Yet these gains are accompanied by significant relational losses that must be critically examined. Unlike professional helping relationships, chatbot-mediated interactions lack authentic reciprocity. Interpersonal trust develops through interaction, shared definition of the situation, and the ongoing negotiation of meaning within a relationship characterised by vulnerability and accountability (as mentioned above). In chatbot conversations, these processes are simulated; trust becomes anchored in the subjective experience of interaction and in the emotional resonance generated by language. Although conversational AI systems may learn from clients' inputs, such learning is oriented toward optimisation, prediction, and data extraction, rather than toward ethical responsibility, relational accountability, or emancipatory aims. In this sense, what appears as learning or adaptation at the level of the system remains external to the ethical horizon of the profession. The user may disclose, but the system does not respond from a position of ethical responsibility; it cannot be questioned or held accountable in the way a professional can. This absence of reciprocity fundamentally alters the moral texture of trust, transforming it from a relational achievement into a functional orientation toward a system.

This reconfiguration of trust-building processes raises pressing ethical and political questions. The very features that make chatbots appealing – control, predictability, and non-judgement – also risk normalising forms of engagement that bypass dialogue, contestation, and collective meaning-making. Chatbots tend to construct a truncated form of relationality that is decontextualised. If help becomes something that is accessed privately, asynchronously, and without relational negotiation, the social and political dimensions of social work interventions risk being further challenged, further reinforcing processes of individualisations. For clients, then, the turn to chatbots may be primarily motivated by the expectation of a less exposing and more controllable forms of engagement, reinforcing instrumental relations with professionals and services also on their sides.

For social workers, structural problems may be reframed as personal issues to be managed through interaction with a system, rather than as collective concerns requiring relational and institutional accountability.

Social work's ethical distinctiveness lies precisely in its capacity to work within, and against, asymmetrical power relations through reflexive practice, making institutional constraints transparent, negotiating aims, recognising clients' experiential knowledge, and exercising discretion responsibly in ways that minimise oppressive effects while promoting self-determination (Sanfelici, 2024). If relational work is replaced – or even systematically marginalised – by automated interfaces, social work risks losing one of its core processes for addressing power imbalances and fostering recognition and trust as a collective effort.

Conclusion

Ultimately, the introduction of AI chatbots into social work practice has been reconfiguring the basis of trust and the very conditions under which help is sought, experienced, and recognised. Trust is no longer grounded exclusively in

professional–client relationships, but increasingly emerges within socio-technical assemblages in which technological artefacts, institutional arrangements, and professional practices are intertwined and governed by partially different ethical and relational logics. In these settings, trust in chatbots is predominantly constructed through interactional cues and algorithmic calculation, rather than through moral agency, normative accountability, or the kind of reciprocal meaning-making that characterises professional helping relationships (Ng and Zhang, 2025; Becker and Fischer, 2024). The appeal of conversational AI therefore rests on a form of relational plausibility: the system can be experienced as listening, understanding, and "being there," thereby lowering thresholds to engagement and offering a sense of recognition, while simultaneously concealing – or at least backgrounding – the absence of ethical responsibility and answerability. This is not a merely technical issue, but a relational and political one, because it changes what service users may come to expect from help: trust becomes anchored in the subjective resonance of interaction and the immediacy of response, rather than in the possibility of negotiating aims, contesting interpretations, and holding an accountable other responsible within a relationship.

This chapter has argued that such a transformation is inseparable from broader tensions concerning the purpose of social work and the models of welfare governance within which practice is situated. The integration of chatbots appears more easily aligned with service delivery–oriented conceptions of social work – centred on efficiency, standardisation, scalability, and individualised problem management – than with social justice-oriented conceptions, in which relational work, professional discretion, and the capacity to locate individual troubles within structural conditions remain central. This distinction is not merely ideological: it shapes which forms of practice organisations promote, which roles practitioners are enabled or constrained to enact, and which modes of support and recognition service users are encouraged to seek and experience. As welfare institutions increasingly adopt logics associated with new public management, digital systems are often justified as tools for streamlining access, reducing costs, ensuring procedural consistency, and minimising discretionary variation. In this context, chatbots may be framed as relatively unproblematic innovations that improve accessibility while containing organisational uncertainty. However, the very conditions that make chatbots attractive can also normalise a truncated relationality in which engagement is decoupled from reciprocity, co-construction, and accountability. The question is not whether chatbots can produce useful interactions, but whether their adoption reinforces a welfare orientation that shifts responsibility for distress toward individual coping and adaptation, while unrecognising and obscuring the value of relational and collective processes of meaning-making.

At the same time, it would be reductive to interpret the turn to chatbots as evidence of user naivety or misplaced trust. Rather, it often represents a response to welfare encounters represented or experienced as difficult to navigate, emotionally risky, and symbolically violent. The popularity of chatbots should therefore be read as both symptom and catalyst: symptom of structural conditions that have eroded accessibility and relational safety in human-centred services, and catalyst for unveiling human-made injustices and redefining of what counts as care,

support, and relationship in contemporary welfare systems. The task for critical social work is not simply to accept or reject conversational AI, but to specify the conditions under which these technologies might be integrated without further hollowing out the moral and relational foundations of help. This requires organisational accountability and governance (including transparency about data use and limits of systems), professional reflexivity about the redistribution of responsibility within hybrid human-AI configurations, and a renewed commitment to relational practice as a site where trust is not merely elicited but ethically earned. Only by holding together the gains that users experience and the relational losses that may follow can social work develop a critical, justice-oriented response to conversational AI that neither romanticises human care nor naturalises automated support as an inevitable future.

References

Afroogh, S., Akbari, A., & Malone, E. et al. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications, 11*, 1568. https://doi.org/10.1057/s41599-024-04044-8

Becker, C., & Fischer, M. (2024). Factors of trust building in conversational AI systems: A literature review. In *5th International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference*, HCII 2024, Washington, DC, June 29–July 4, 2024, Part II.

Behnia, B. (2008). Trust development: A discussion of three approaches and a proposed alternative. *The British Journal of Social Work*, *38*(7), 1425–1441. https://doi.org/10.1093/bjsw/bcm053

Betlemidze, M. (2022). Traversing anthropocentric horizons with her: Trans-corporeal surrogacy, enchantment, and disenchantment in human-machine assemblage. *Journal of Communication Inquiry*, *46*(2), 206–224.

Blomqvist, K., & Stahle, P. (2004). Trust in technology partnership. In M. L. Huotri & M. I. Hershey (Eds.), *Trust in knowledge management and systems in organizations* (pp. 173–199). Idea Group Publishing.

Bos, N., Olson, J. S., Gergle, D., & Olson, G. M. (2002). Effects of four computer-mediated communications channels on trust development. In *Proceedings of the CHI 2002 Conference on Human Factors in Computing Systems: Changing our World, Changing Ourselves*, Minneapolis, MN, April 20–25.

Chan, C., & Li, F. (2023). Developing a natural language-based AI-chatbot for social work training: An illustrative case study. *China Journal of Social Work*, *16*(2), 121–136. https://doi.org/10.1080/17525098.2023.2176901

Eyal, G., Au, L., & Capotescu, C. (2024). Trust is a verb!: A critical reconstruction of the sociological theory of trust. *Sociologica*, *18*(2), 169–191. https://doi.org/10.6092/issn.1971-8853/19316

Grassi, E. (2025). Emotional intersections: Exploring the ontogenesis of human-algorithm relationships. *Sociologia Italiana*, *29*, 11–34. https://doi.org/10.82031/2281-2652-202529-1.

Greer, S., Ramo, D., Chang, Y. J., Fu, M., Moskowitz, J., & Haritatos, J. (2019). Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomised controlled feasibility trial. *JMIR mHealth and uHealth, 7*(10), Article e15018. https://doi.org/10.2196/15018.

Gursoy, D., Chi, O. H., Lu, L., & Nunkoo, R. (2019). Consumers acceptance of artificially intelligent (AI) device use in service delivery. *International Journal of Information Management*, *49*, 157–169. https://doi.org/10.1016/j.ijinfomgt.2019.03.008

He, L., Basar, E., Wiers, R. W., Antheunis, M. L., & Krahmer, E. (2022). Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*, *220*, Article 726. https://doi.org/10.1186/s12889-022-13115-x

Hepp, A., Hjarvard, S., & Lundby, K. (2015). Mediatisation: Theorizing the interplay between media, culture and society. *Media Culture & Society*, *37*(2), 314–324. https://doi.org/10.1177/0163443715573835

Khawaja, Z., & Bélisle-Pipon J-C. (2023). Your robot therapist is not your therapist: Understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health, 5*, 1278186. https://doi.org/10.3389/fdgth.2023.1278186

Khodyakov, D. (2007). Trust as a process: A three-dimensional approach: A three-dimensional approach. *Sociology*, *41*(1), 115–132. https://doi.org/10.1177/0038038507072285

Levin, D. Z., Whitener, E. M., & Cross, R. (2006) Perceived trustworthiness of knowledge sources: The moderating impact of relationship length, *Journal of Applied Psychology*, *91*(5), 1163–1187. https://doi.org/10.1037/0021-9010.91.5.1163

Lipsky, M. (2010). *Street-level bureaucracy: Dilemmas of the individual in public services*. Russell Sage.

Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomised trial of effectiveness. *Internet Interventions*, *27*, 100495. https://doi.org/10.1016/j.invent.2022.100495

Luhmann, N. (1979). *Trust and power*. John Wiley & Sons.

Ng, S. W. T, & Zhang, R. (2025). Trust in AI chatbots: A systematic review. *Telematics and Informatics*, *97*, 102240. https://doi.org/10.1016/j.tele.2025.102240

Northhouse, P. G., & Northhouse, L. L. (1985). *Health communication: A handbook for health professionals*. Prentice-Hall.

Orlikowsky, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organisations. *Organisation Science*, *11*(4), 404–428. https://www.jstor.org/stable/2640412

Oshni Alvandi, E. (2019). Cybertherapogy: A conceptual architecting of presence for counselling via technology. *International Journal of Psychology and Educational Studies*, *6*(1), 30–45. https://doi.org/10.17220/ijpes.2019.01.004

Pellegrino, E. D. (1991). Trust and distrust in professional ethics. In E. D. Pellegrino, R. M. Veatch & J. P. Langan (Eds.), *Ethics, trust, and the professions: Philosophi cal and cultural aspects* (pp, 69–89). Georgetown University Press.

Ruch, G., Turney, D., & Ward, A. (2018). *Relationship-based social work: Getting to the heart of practice*. Jessica Kingsley.

Sanfelici, M. (2024). *Fondamenti del servizio sociale anti-oppressivo*. Carocci.

Sanfelici, M., & Nothdurfter, U. (2026) (eds.). *La digitalizzazione del lavoro sociale. Studi di caso nei servizi per la salute.* Carocci.

Trevithick, P. (2003). Effective relationship-based practice: A theoretical exploration. *Journal of Social Work Practice*, *17*(2), 163–176. https://doi.org/10.1080/026505302000145699

Victor, B. G., & Goldkind, L. (2025) The therapist in the machine: Confronting AI's challenge to clinical social work. *Journal of Technology in Human Services*, *43*(2), 73–81. https://doi.org/10.1080/15228835.2025.2500827

11 Care, Artificial Intelligence (AI), and Ethics

New Opportunities and Challenges for Social Workers

Eleni Papouli

Introduction

Social work has historically been anthropocentric, centred on the well-being of humans and communities. However, as we delve deeper into the nuances of care within social work, especially in light of emerging advancements in Artificial Intelligence (AI) and post-anthropocentric philosophy, we find ourselves at a critical juncture. The meaning of care is evolving, prompting a reevaluation of its implications for social workers in a landscape increasingly dominated by technology. As society incorporates AI, from virtual platforms to predictive systems for caring humans and nonhuman beings and ecosystems, it is crucial to ask whether traditional principles of care and their ethical foundations can retain significance. This raises critical discussions about the opportunities and challenges social workers face in contexts where care involves not only human needs but also the well-being of animals, ecosystems, and the planet. In post-anthropocentric contexts where humans, animals, and technologically advanced entities may interact, how can social workers maintain their commitment to promoting well-being and addressing systemic inequalities that affect not only humanity but also animals and ecosystems? For social work, this means recognising that "care" involves challenging policies and practices that perpetuate these interlinked injustices, whether by advocating for environmental protections, supporting multispecies well-being, or addressing the ecological costs of AI systems themselves.

In this chapter, I adopt a post-anthropocentric perspective that emphasises the interdependence of humans, nonhuman species, and ecosystems, advocating care practices that recognise the agency of nonhuman life (Bell, 2023; Bozalek & Zembylas, 2025). Through this lens, my analysis foregrounds multispecies interconnectedness and ecological responsibility, while engaging posthumanist thought to interrogate the limitations of human-centred ethics. Although posthumanism is sometimes used interchangeably with post-anthropocentrism in the literature, it represents a distinct perspective. Posthumanism extends the critique further, challenging the very centrality of the human subject and, in some formulations, conceptualising AI and robots as participants in posthuman worlds (Braidotti, 2013; Puig de la Bellacasa, 2017). I also draw on Tronto's framework of care (Tronto, 2013, 2015), extending its political dimensions into post-anthropocentric debates to explore how social work can engage

DOI: 10.4324/9781003627920-11
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

with care involving AI, ecosystems, and nonhuman entities. This extension represents my interpretive contribution, exploring how Tronto's ideas about the political dimensions of care can illuminate emerging contexts, rather than asserting her personal stance. The chapter critically examines both the promises and perils of AI in care, emphasising the need for social workers to engage with an AI-enhanced future while attending to ecological, ethical, and political responsibilities.

Care as Sociopolitical Praxis: An Overview

Care, whether professional or personal, is central to human and social life. Barnes (2012) argues that care is a universal human need, but it is also present in nonhuman worlds, where many species nurture offspring. In human societies, care is closely tied to welfare systems, sparking debates about state responsibility and social justice. Since the 1960s, feminist movements have reframed care as an ethical and political issue, with Carol Gilligan (2016), who is considered the originator of the ethics of care, linking care to gender and inequality. Her work challenged patriarchal knowledge structures, framing care as transformative.

In the 1990s, American feminist political scientist Joan Tronto, a leading figure in care ethics, advanced the discourse by articulating the political dimensions of care. She argued that care must lie at the heart of democratic life: all policies, not only social policies, should be evaluated by how well they sustain caring relationships and equitable interdependence. Tronto identified phases of care–caring about, caring for, caregiving, and care receiving– and later added caring with, emphasising collective responsibility, solidarity, trust, and fairness (Tronto, 2013, 2015). Her framework highlights the complexities and systemic dimensions of care, revealing its critical role in social justice and community resilience. For Tronto, caring offers transformative political insights for addressing systemic injustices and institutional violence while strengthening collective well-being.

While Tronto's work is primarily human-centred and does not explicitly address care practices involving the natural environment or technologies such as AI and robots – as mediators of care, care agents, or objects in care relations–her concepts of caring for and caring with reflect concern for broader collectives beyond the immediate human circle (Parra Jounou & Tronto, 2024). Viewed through a post-anthropocentric lens, these concepts resonate with principles of collective responsibility, intergenerational justice, and ecological sustainability. They encourage practitioners and scholars to consider how care responsibilities are allocated and shared across communities, generations, and even species and ecosystems, providing guidance for social work in addressing emerging ethical challenges posed by AI and other technological interventions.

Care in Social Work: An Ethical Imperative

Care is the ethical cornerstone of social work, yet its meaning is complex and multifaceted. It encompasses both the relationships between social workers and clients and the state's responsibility to provide resources. This dual lens highlights

care's intricate nature, bridging personal connections and systemic obligations in the pursuit of social welfare. Tronto (2013, 2015) emphasises that care entails both relational ethics and political responsibility to deliver dignified interventions and community support. Practicing care builds trust and fosters relationships, central to effective, person-centred social work. It acknowledges the vulnerability of those receiving support while extending responsibility to others, the wider community, and society as a whole.

Historically human-centred social work focuses on supporting people. Yet, as practitioners navigate increasingly complex dynamics including interactions with other species, AI technologies, and the environment, the principles of care must evolve to meet these emerging ethical and professional challenges. A post-anthropocentric ethic of care challenges the profession to embrace interdependence, sustainability, and inclusivity, ensuring that social work practice fosters flourishing across both human and nonhuman communities (Bell, 2023; Bozalek & Zembylas, 2025; Papouli, Syropoulou & Kyriazi, 2025). In this context, social work's role as a caregiver expands significantly beyond its conventional human-centred scope in a post-anthropocentric world, where human-centred perspectives are no longer dominant. It encompasses not only connections between social workers and humans but also responsibilities towards nonhuman life, ecosystems, and the broader socio-ecological systems that sustain well-being. This necessitates a redefinition of care to encompass multispecies and ecological welfare, promoting relational ethics, and actively confronting speciesism and other forms of oppression that transcend human interactions (Bell, 2023; Cañada, Butcher & Sariola, 2022; Papouli, Syropoulou & Kyriazi, 2025; Silberberg, 2023). As will be seen later, by adopting a post-anthropocentric view of care, social workers can help create a more equitable social landscape that respects the rights and needs of both humans and nonhuman entities, ultimately fostering a more inclusive planet for all.

Caring for the More-than-Human World

Today, progressive social work scholars call for a broader understanding of care that extends beyond human interests. This view is reflected in ecosocial approaches to social work (IFSW, 2022) and in feminist – particularly ecofeminist – ethics of care, which critique neoliberal capitalism's role in environmental harm, highlight the interconnectedness of all life, and promote a holistic conception of well-being (Papouli, Syropoulou & Kyriazi, 2025). As discussed, when we move beyond a human-centred view and recognise how closely connected humans and non-humans are, care can be seen as extending to animals and ecosystems. In technology-mediated relationships, including those involving AI, notably, a posthuman perspective frames AI not just as a tool but as part of shared care practices (Braidotti, 2013; Puig de la Bellacasa, 2017). As Braidotti (2013) argues, this view challenges human-centred assumptions and highlights the interdependence of humans, animals, and even machines. In this light, caring with and caring for move beyond exclusively human relationships to embrace all forms of existence, encouraging social workers to consider how AI technologies might contribute to caregiving and protection (Coeckelbergh, 2021; Johnston &

Taylor, 2020). Although there is much debate about whether machines could one day become sentient, today's AI systems merely simulate intelligent behaviour and do not possess independent moral agency.

Aristotle argued that human beings' true nature, virtue, and flourishing cannot be realised in isolation but only within a political community (Papouli, 2019). Being is never an isolated or self-sufficient entity; rather, it emerges relationally, co-constituted through entanglements with other humans, nonhuman animals, and ecological systems. Recognising this ontological interdependence calls for extending ethical concern beyond the human, fostering care practices and politics that sustain more just, inclusive, and ecologically sustainable communities (Bozalek & Zembylas, 2025).

Notably, caring for the more-than-human world challenges human exceptionalism without, however, diminishing the value of human life. Nor does it justify harming people in the name of protecting the environment. Instead, it expands our ethical responsibility to include the beings and systems on which our own well-being depends. This perspective affirms the intrinsic value of all life, including human life, and calls for democratic, nonviolent approaches to ecological care. Within this vision, social work, rooted in human rights and social justice, cannot endorse any agenda, sometimes called "eco-fascism" (Campion, 2021; Zimmerman, 1995) that sacrifices vulnerable people for supposed environmental goals. Rather, post-anthropocentric social work seeks to transform social, economic, and technological systems so that human and nonhuman flourishing are mutually reinforcing.

AI in the Service of Care: Opportunities for Social Workers

The use of digital technology in social work dates back to the early 1980s, so the interest in this area is not new (Reamer, 2013). Nonetheless, it was during the pandemic that its use surged, sparking interest in the possibility of incorporating AI technologies into social work to improve service delivery methods, particularly in developed countries. This marked a transformative shift in how professionals approached various challenges in the field. Initially, early AI applications focused on predictive analytics, which were used to assess and predict risks and outcomes in various areas, such as mental health and child welfare and family services (Reamer, 2023). Since 2020, there has been a significant increase in scholarly literature and research focusing on the diverse roles of AI in social work settings.

AI is increasingly recognised as a care technology capable of complementing and transforming human-centred support systems while extending care to multispecies and ecological contexts. AI is being integrated across diverse care practices, from predictive analytics in social services and algorithmic triage in health and crisis response to initiatives supporting environmental protection and disaster management (IPCC, 2023; UNEP, 2024). Here, I discuss various applications of AI, including its role in delivering targeted and personalised interventions for care recipients.

Predictive Care

AI-powered predictive care is reshaping social work by using machine learning and decision-support systems to anticipate crises such as homelessness or mental health emergencies before they occur (Reamer, 2023; Sabry et al., 2022). Data from health records, wearable devices, and environmental sensors allow early detection of risk patterns and enable tailored interventions. A post-anthropocentric perspective expands this model beyond a strictly human focus. It urges recognition that well-being is interdependent with ecosystems and other species, and that predictive systems should account for environmental factors like climate stressors and pollution. By tailoring interventions to both human and nonhuman factors, social workers are potentially empowered to deliver care that is more responsive to shared ecologies. This includes recognising that human well-being is often deeply interdependent with the health of ecosystems and other species, and hence implementing predictive care in ways that are ethically sustainable in more-than-human worlds.

Chatbots, Robotic Systems and Virtual Assistants

From a post-anthropocentric, particularly posthuman, standpoint, chatbots, robotic systems, and virtual assistants may be understood as constitutive participants within a heterogeneous socio-technical assemblage, co-producing modalities of care that extend beyond conventional human-to-human relational frameworks. Their engagements with users – whether through the provision of guidance, mediation of information, or facilitation of affective attunement – instantiate a distributed and relational form of care, wherein agency is enacted collectively by human actors, technological entities, and the material and ecological contexts they inhabit. Within the domain of social work, this perspective invites practitioners to reconceptualise virtual companions not merely as instruments of service delivery but as co-constitutive agents in an emergent assemblage of well-being, encompassing technological, ecological, and human dimensions.

Management Practices

AI tools used for administrative and managerial tasks can enhance the efficiency and effectiveness of social work while enabling a broader, post-anthropocentric understanding of care. By automating time-consuming and repetitive activities such as appointment scheduling, documentation, note-taking, and case tracking, these technologies allow social workers to devote more attention to meaningful interactions (Reamer, 2023), not only with human clients but also with the broader networks of nonhuman and ecological systems that influence well-being. This shift supports more nuanced decision-making and holistic care, acknowledging that human welfare is deeply intertwined with the health of communities, ecosystems, and other species. In this sense, AI does more than optimise workflows; it opens the possibility of reimagining care as a relational, multispecies practice, attentive

to interconnected lives and environments (Pineda-Pinto, 2025; Rakova & Dobbe, 2023). Nonetheless, the role of AI in decision-making is a critical concern; as Segal (2024) points out, AI system-generated responses may lack the subtle understanding required to make complex ethical decisions in social work.

Environmental Sustainability

Environmental care is not an optional add-on but a core ethical practice (Bell, 2023). Although the proliferation of AI is increasing energy consumption and raising concerns about efforts to achieve net-zero targets, as discussed in the next section, social workers can help protect shared ecosystems and mitigate climate change by collaborating inter-disciplinarily with other professionals and by utilising AI-driven tools such as predictive climate-risk mapping and sensor networks that monitor soil and water quality. For instance, practitioners might work with coastal communities and marine biologists to design relocation plans that safeguard human livelihoods while respecting the migratory patterns of marine species. AI can also assist communities, individuals, and their animals affected by natural disasters by facilitating access to resources, emergency aid, housing, and mental health services (Eichler, 2025). Bell (2023) emphasises that by viewing animals, rivers, and forests as partners rather than mere resources, post-anthropocentric social work fosters planetary sustainability and protects the most vulnerable, both human and more-than-human.

AI, Care and Social Work: Confronting the Emerging Challenges

The rapid emergence of AI technologies is reshaping how social workers interact with care recipients and introducing a range of challenges, particularly concerning the ethical use of AI in practice. In response, scholars have begun examining AI's complex implications from both security and ethical perspectives. A review of the literature in social work and related disciplines identifies several key areas of ethical concern, which are presented and analysed below. Because ethical questions in social work are inherently political, the role of AI in shaping the politics of care remains an integral part of this discussion.

Impact on Care Recipients

Recent literature is beginning to unsettle the assumption that AI-enabled care exists solely for the benefit – or risk – of human recipients. Rather than treating "quality of care," "client autonomy," or "privacy" as exclusively human concerns, a post-anthropocentric perspective recognises that algorithmic infrastructures reconfigure relations among people, nonhuman organisms, and the material environment. Building on this expanded view, analyses of algorithmic bias (Afrouz & Lucas, 2023; Barsky, 2023; Berridge, Goldkind & Bricout, 2024; Reamer, 2023) can be understood not only as threats to equitable human service but also as forces that redistribute vulnerability across species and ecosystems. Hoche et al. (2025) highlight machine learning's structural tendency toward patterned bias. From a

more-than-human standpoint, such bias is not just a statistical flaw but a mechanism that privileges certain landscapes, languages, and life-forms while erasing others. For example, predictive systems may inherit biases not only among human populations (race, gender, and class) but also biases of speciesism or neglect of nonhuman suffering. AI fairness frameworks tend to overlook harms to animals or ecosystems, even though technologies can impact nonhuman well-being substantially (IPCC, 2023; UNEP, 2024).

Refocusing on human-centred care, the use of wearables and related technologies also introduces significant ethical concerns, such as surveillance, privacy, data ownership, and the risk of dehumanising care – or what some scholars (Canali et al., 2024) describe as the "datafication" of care relationships –where individuals are treated as mere nodes within predictive systems. These challenges are especially pronounced in areas like healthy ageing, where questions of justice, autonomy, and inclusion intersect closely with technological interventions (Canali et al., 2024).

Based on the literature review, AI systems appear to pose many of the same ethical risks that have emerged over the past few decades with the Internet and technologies used in social services, particularly regarding data collection, privacy, confidentiality, storage, and access (Papouli, 2020). The key distinction with AI lies in the scale of data collection and the automated processing capabilities driven by algorithms. In the context of automated AI care provision, social workers have expressed concerns about how to evaluate the quality of care delivered by robots compared to trained human professionals. Additionally, questions arise about what happens when a care recipient perceives that a robot is not providing adequate support but is compelled to use it due to institutional policies or financial constraints. There are also concerns regarding the use of AI to convey empathy and the risks of overreliance on empathetic AI. The recent death[1] of a 14-year-old, and the subsequent lawsuit alleging that an AI chatbot influenced him to commit suicide, underscores the serious ethical concerns surrounding the psychological impact and potential risks of human-AI interactions. Consequently, debates about the "quality of care" shift in meaning when robots and algorithmic agents are considered co-actors rather than mere tools.

As AI tools are increasingly used in personal and emotionally charged contexts, governments, alongside and often primarily, social movements and independent organisations, have pressured AI companies to revise their policies and implement comprehensive ethical and operational reforms to prevent harm and discrimination. In response to this political and social pressure, AI companies using OpenAI models, such as ChatGPT, have recently announced measures to identify and filter situations that may cause psychological stress, thereby assisting in the prevention of responses that could exacerbate, for example, delusional thinking or emotional dependence (LIFO, 2025).

Caring Professions in the Age of AI

The literature highlights the evolving landscape of health and caring professions, including the social work profession as a result of AI integration; it delves into the changing nature of human service professions, looking at how AI technologies are

reshaping practitioners' roles and responsibilities, as well as the potential to redefine professional relationships and care delivery. This shift raises questions about the roles and responsibilities of care providers, the necessity for new skill sets, and the potential for loss of human connection and job displacement. Human connection and the ability to engage with clients in their most vulnerable moments are at the heart of social work's caring practice. By replacing these with AI tools that become more advanced and support automated tasks, the profession risks being dehumanised and social workers being replaced, reducing the need for the workforce. The AI boom has sparked a heated debate about how AI can supplement, rather than replace, human services and professionals, as well as concerns about the potential decline in critical thinking skills. Scholars (e.g., Barsky, 2024; Coulthard, Taylor & McGlade, 2025; Reamer, 2023) emphasise the importance of balancing human care and AI technologies through a collaborative practice approach, in which social workers, care recipients, community representatives, and technologists jointly design and monitor AI tools to ensure they reinforce, rather than erode, ethical and relational aspects of care.

Disparities and Justice-Related Issues

How can we make sure that everyone has fair access to advanced caregiving technologies while also caring for the well-being of humans, other species, and the planet? This question reminds us that equity and justice shouldn't stop at people alone. AI in health and care can help improve lives, but it can also deepen social inequalities and reflect human-centred biases. From a more-than-human-world perspective, fairness means more than just giving people access to new AI tools. It also means protecting the environments and species affected by the energy use, resource extraction, and waste these technologies create.

In the current global context, structural inequalities continue to intensify existing digital disparities, thereby deepening the emerging "AI divide" (ILO, 2024). This divide is particularly salient in low-income regions, where complex socio-economic, infrastructural, and educational constraints continue to impede equitable access to digital technologies and services (Mastrothanasis et al., 2025; WHO, 2024). Moreover, AI systems often overlook diverse cultural contexts and rarely account for the intrinsic value of nonhuman life. Regarding values, humans hold conflicting gender-related, ecological, and sociopolitical perspectives, ranging from democratic to authoritarian or fundamentalist orientations. Which of these values should guide the development of AI tools aimed at addressing human care needs? Additionally, whose priorities should shape AI design when care must extend to both human and more-than-human communities?

Given that AI tools can reproduce social biases and obscure accountability, scholars stress the importance of continuous critical scrutiny to prevent social injustices (Afrouz & Lucas, 2023; Barsky, 2024; Reamer, 2023), mitigate environmental harm, and promote fairness throughout the web of life.

AI's Effect on the Environment

AI systems have significant ecological impacts, including substantial energy consumption, rising emissions, electronic waste, and heavy water usage (UNEP, 2024; Zewe, 2025). From a post-anthropocentric social work perspective, these issues are not only as environmental concerns but also matters of social justice, as vulnerable populations and marginalised communities are disproportionately affected. By bridging technology, the environment, and human rights, social workers can help ensure that the benefits of AI do not come at the expense of ecological health or social equity. Through advocacy, education, and collaboration, they promote AI-mediated care practices that are both socially responsible and environmentally sustainable.

AI and the Politics of Care

What matters for social work today is how AI tools are already reshaping the politics of care: shifting authority from practitioners to proprietary algorithms, influencing which needs are recognised or ignored through data selection and modelling, and introducing new forms of surveillance and accountability that can either reinforce or challenge social inequities. Focusing on these tangible dynamics enables social workers to address AI as a political and ethical issue now, rather than speculating about distant possibilities of machine consciousness. Following the earlier analysis, AI can be viewed both as a tool of care and as a site of political contestation. This dual role challenges the claim that AI makes care apolitical; instead, it reshapes the political dynamics in which care unfolds. Choices about which needs justify data collection, how populations are classified, which nonhuman entities are protected from harm, and how automated systems are governed are themselves acts of political power.

Preparing Social Workers for an AI-enhanced Future of Care

As societies transition into a post-anthropocentric era shaped by advanced technologies, the social work profession must reconceptualise care beyond the human, extending ethical and relational responsibilities to nonhuman entities and ecosystems. In this context, AI should be critically engaged not only as a tool for facilitating care but also as an integral component of the complex, interdependent web of planetary life – both shaping and being shaped by the broader ecological and social systems within which care is enacted.

At present, countries around the world are developing or have already established national standards for the responsible use of AI in public and private health and social care services, aiming to address ethical and legal concerns (Crocker, 2025; Reamer, 2023). However, knowledge remains limited regarding the specific legislation in place and how these regulations are implemented, as research on this topic is still scarce. Within social work, critical discussion of AI safety and risk has already prompted international and national associations to revise and strengthen their ethical codes, both to address potential threats and to engage with possible solutions. For example, recent NASW (2023) and BASW (2025) statements on

AI ethics suggest that the profession has begun to engage with the governance of emerging technologies. I reference these not as models to follow uncritically, but as indicators of the current professional baseline, a starting point for envisioning more expansive frameworks that embrace ecological interdependence and multispecies justice. These documents, however, remain rooted in a human-centred understanding of care and do not yet incorporate post-anthropocentric perspectives.

In light of the chapter's discussion, there is an urgent need for ongoing dialogue and research into the implications of using AI for care in social work, ensuring that the profession's core values remain central to practice. While international social work organisations such as the IFSW (2018) articulate principles like human dignity, the worth of the person, and a commitment to social justice primarily in human terms, a post-anthropocentric perspective calls for a more inclusive expression of these values. In this view, "person" can be understood not solely as a human category but as recognition of the intrinsic value of nonhuman beings and ecosystems. From this standpoint, the profession's foundational values – care, justice, and respect – extend to animals, plant life, and the ecological systems that sustain all species. Ja' Bree Harris (2025) observes that AI is already reshaping sectors such as healthcare and criminal justice, and social work is no exception. In response, she advocates for the establishment of an AI Commission dedicated specifically to the needs and ethics of the social work profession.

Furthermore, the literature on ethical issues about the use of AI in social work demonstrates the value of training both students and practitioners in the opportunities and limitations of AI technologies (Ahn et al., 2025; Mathiyazhagan & Patton, 2025; Reamer, 2023). Such training equips professionals to navigate an emerging "technocaring" landscape that extends beyond traditional human interactions, reflecting the evolving dynamics of posthuman societies. Recognising the importance of AI literacy, Ahn and colleagues (2025) contend that social workers must strengthen their AI skills to (a) understand and help mitigate the social inequalities AI can exacerbate, (b) guide the ethical development and management of AI systems, and (c) use AI tools judiciously in practice (p. 8). Social workers should approach AI as a contributor to collective flourishing rather than a universal remedy. This requires critical vigilance regarding the interplay of ethical principles, legal norms, and planetary limits, and calls for the development of technocare ethics attuned to post-anthropocentric societies. While AI may support care practices, it must remain a companion to, never a replacement for, the situated wisdom of social workers (Barsky, 2023).

Beyond the need for training, some scholars argue that social workers with technological expertise should participate in AI development to help protect vulnerable and marginalised groups from harm or discrimination. Berridge et al. (2024) note that while many technology leaders highlight AI's potential benefits, social workers offer crucial insight for addressing current algorithmic harms and ensuring ethical, responsible use. The authors suggest that this involvement can be guided by key questions such as: (1) Participation: How were users and affected communities involved in the system's design? (2) Representation: How were their experiences incorporated, and who ultimately benefits from automation? (3) Data and assumptions: What texts and

data sets trained the model, what or who is excluded, and which cultural assumptions shape its analysis? Understanding the training data and cultural perspectives embedded in AI tools is thus critical for ensuring equity and justice.

Building on earlier points, integrating AI tools into social work often frames care as a problem of statistical optimisation, identifying risks, predicting outcomes, and allocating resources. While such tools can improve efficiency, they may also obscure the inherently political nature of care that Tronto emphasises. Decisions about which data to collect, which outcomes to prioritise, and how algorithms balance competing needs are themselves political acts, yet they are frequently presented as neutral or value-free. Consequently, the use of AI as a care technology in social work cannot be regarded as a mere technical afterthought. If care is understood as a political principle, the design and implementation of AI systems must be democratic and participatory, prioritising relational ethics and social justice. From a post-anthropocentric perspective, incorporating AI into social work calls for a multispecies ethics of care – one that upholds human dignity, promotes ecological sustainability, and recognises the agency of nonhuman beings – to ensure that predictive innovations support rather than undermine shared ecologies of care. In this context, questions of dataset selection, algorithmic transparency, and accountability are more than engineering concerns; they embody the collective responsibility that Tronto (2013) identifies as central to a caring democracy.

AI technologies are rapidly evolving and becoming integral to our post-anthropocentric world, shifting the meaning and scope of care and prompting social work research into AI's global impact on humans, nonhuman beings, and ecosystems. To imagine futures where technology, multispecies care, and the living earth co-evolve, social work must move beyond human exceptionalism. Best practices should empower practitioners to advocate for the well-being of intertwined human and more-than-human communities, while honouring foundational professional values such as empathy, solidarity, and justice across species and ecosystems. Understanding these shifting dynamics equips social workers to engage with a world in which the very meaning of care expands to encompass reciprocal relationships among people, other beings, and the environments that sustain them. Through such exploration, the field can grow not only resilient but also inventive, fulfilling its mission within complex ecological and technological realities.

Conclusion

The rise of AI technologies has sparked new discussions about how we care for one another and for our environment – understood as the more-than-human world. Caring for this environment means recognising the intrinsic value and agency of nonhuman beings and ecosystems, and shaping technology including AI, so that it supports the flourishing of this wider community of life. Yet AI's promise of optimisation can obscure the political contestation and collective responsibility that Tronto deems essential to democratic care. My argument, therefore, does not advocate a technocratic replacement of care; rather, it calls for vigilance to ensure that AI remains accountable to social-justice aims.

AI in care practices is more than a mere tool. It is an ethical, ecological, and political force that warrants open public debate. Recognising that algorithmic interventions can both reveal and obscure the politics of care, practitioners can critically engage with these systems to ensure that care remains grounded in social work's core commitments while extending moral regard to nonhuman beings and the planetary systems that sustain life. By embedding AI within interconnected ethical, ecological, and political discourses, social workers can foster holistic, equitable, and compassionate forms of care that honour the vitality of the entire earth community.

Note

1 Can A.I. be blamed for a teen's suicide? https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html.

References

Afrouz, R., & Lucas, J. (2023). A systematic review of technology-mediated social work practice: Benefits, uncertainties, and future directions. *Journal of Social Work*, *23*(5), 953–974. https://doi.org/10.1177/14680173231165926

Ahn, E., Choi, M., & Fowler, P., & Song, I. H. (2025). Artificial Intelligence (AI) literacy for social work: Implications for core competencies. *Journal of the Society for Social Work and Research (Ahead of print)*. https://doi.org/10.1086/735187

Barnes, M. (2012). *Care in everyday life: An ethic of care in practice*. Policy Press.

Barsky, A. E. (2024). *The ethics of artificial intelligence (AI) in social work practice*. [PowerPoint slides]. https://barsky.org/

Barsky, A. E. (2023, July 7). Ethics alive! Artificial intelligence: Tricks or tools for social work education and practice. *The New Social Worker*. https://www.socialworker.com/feature-articles/ethics-articles/artificial-intelligence-tricks-or-tools-for-social-work-education-practice

BASW. (2025). AI, social work, and our shared future: A BASW UK perspective. Retrieved March 7, 2025. https://new.basw.co.uk/articles/ai-social-work-and-our-shared-future-basw-uk-perspective?fbclid=IwY2xjawJ9o2pleHRuA2FlbQIxMQABHntO1SzlhGVO4WMDxygpjhvGXwSeT0PNZ50w8wyvFiSkD5h5vkCmXvtpuc93_aem_EHJoIb6AD7Ehdu6pNZYuXg&sfnsn=mo

Bell, K. (2023). Post-anthropocentric social work. In D. Hölscher, R. Hugman & D. McAuliffe (Eds.), *Social work theory and ethics. Social work* (pp. 121–139). Springer.

Berridge, C., Goldkind, L., & Bricout, J. (2024). If AI has you feeling clueless and concerned, you're in good company. *Social Work Advocates Magazine*. National Association of Social Workers.

Bozalek, V., & Zembylas, M. (2025). Sensibilities of the ocean world: Toward a more-than-human ethics in social work. *Ethics and Social Welfare*, 1–18. https://doi.org/10.1080/17496535.2025.2551301

Braidotti, R. (2013) *The posthuman*. Polity Press.

Campion, K. (2021). Defining ecofascism: Historical foundations and contemporary interpretations in the extreme right. *Terrorism and Political Violence*, *35*(4), 926–944. https://doi.org/10.1080/09546553.2021.1987895

Cañada, J. A., Butcher, A., & Sariola, S. (2022). Beyond anthropocentrism: A relational, ambivalent and care ethics framework for antimicrobial resistance. *Medical Humanities*, *48*(4), 493–501. https://doi.org/10.1136/medhum-2021-012309

Canali, S., Ferretti, A., Schiaffonati, V. et al. (2024). Wearable technologies for healthy ageing: Prospects, challenges, and ethical considerations. *Journal Frailty Aging, 13*, 149–156. https://doi.org/10.14283/jfa.2024.19

Coeckelbergh, M. (2021) *AI ethics*. MIT Press.

Coulthard, B., Taylor, B. J., & McGlade, A. (2025). Artificial intelligence and evidence for social work: will a robot steal your job? *European Social Work Research*. Retrieved May 4, 2025. https://doi.org/10.1332/27551768Y2025D000000033

Crocker, T. (2025). Balancing data privacy and social good: A framework for the AI generation of social and human services. *International Social Work*. https://doi.org/10.1177/00208728251336944

Eichler, A. (2025, July 15). *Using AI to map environmental risks*. New York Institute of Technology. https://www.nyit.edu/news/articles/using-ai-to-map-out-environmental-risks/

Gilligan, C. (2016). *In a different voice: Psychological theory and women's development* (3rd ed.). Harvard University Press.

Hoche, M., Mineeva, O., Rätsch, G., Vayena, E., & Blasimme, A. (2025). What makes clinical machine learning fair? A practical ethics framework. *PLOS Digital Health, 4*(3), e0000728. https://doi.org/10.1371/journal.pdig.0000728

IFSW. (2018). Global social work statement of ethical principles. https://www.ifsw.org/global-social-work-statement-of-ethical-principles/

IFSW. (2022). The role of social workers in advancing a new eco-social world. https://www.ifsw.org/the-role-of-social-workers-in-advancing-a-new-eco-social-world.

Intergovernmental Panel on Climate Change (IPCC). (2023). *Sixth assessment synthesis report: Climate change 2023*. IPCC.

International Labour Organisation (ILO). (2024). Mind the AI divide: Shaping a global perspective on the future of work. *International Labour Organisation*. https://www.ilo.org/publications/major-publications/mind-ai-divide-shaping-global-perspective-future-work

Johnston, A., & Taylor, B. (2020). Artificial intelligence and ethical challenges in social work: Toward post-humanist practice. *Journal of Social Work Values and Ethics*, *17*(2), 45–62.

LIFO. (2025, August 6). *OpenAI adds 'psychological filters' to ChatGPT – Admits it reinforced users' delusions*. https://www.lifo.gr/now/world/i-openai-prosthetei-psyhologika-filtra-sto-chatgpt-paradehetai-oti-enishyse-aytapates

Mastrothanasis, K., Pikoulis, E., Kladaki, M., Pikouli, A., Karamagioli, E., & Papantoniou, D. (2025). Digital drama-based interventions in emergency remote teaching: Enhancing bilingual literacy and psychosocial support during polycrisis. *Psychology International*, *7*(2), 53. https://doi.org/10.3390/psycholint7020053

Mathiyazhagan, S., & Patton, D. U. (2025). Towards just and equitable Web3: Social work recommendations for inclusive practice of AI policies. *AI & Society*, *40*, 1549–1551. https://doi.org/10.1007/s00146-024-01994-0

NASW. (2023). Artificial intelligence in social work: Exploring key ethical considerations. https://www.socialworkers.org/About/Ethics/AI-and-Social-Work

Papouli, E., Syropoulou, D., & Kyriazi, E. (2025). Ecofeminism and social work in the post-anthropocentric era. Common considerations and approaches. *Social Work* (*Social Sciences Review)*, *39*(1), 3–20. https://doi.org/10.12681/socialwork-rss.40649

Papouli, E. (2020). *Social services, social workers and digital technology in the age of COVID-19 pandemic: New moral and ethical challenges*. Retrieved January 20, 2024. https://socialpolicy.gr/2020/04/

Papouli, E. (2019). Aristotle's virtue ethics as a conceptual framework for the study and practice of social work in modern times. *European Journal of Social Work*, *22*(6), 921–934. https://doi.org/10.1080/13691457.2018.1461072

Parra Jounou, I., & Tronto, J. C. (2024). Care ethics in theory and practice: Joan C. Tronto in conversation with Iris Parra Jounou. *Contemporary Political Theory*, *23*(2), 269–283. https://doi.org/10.1057/s41296-024-00680-6

Pineda-Pinto, M. (2025). Realising multispecies justice through a capability approach. *Nature Sustainability, 8*(1), 1–10. https://doi.org/10.1038/s42949-025-00205-z

Puig de la Bellacasa, M. (2017). *Matters of care: Speculative ethics in more than human worlds, Posthumanities*. University of Minnesota Press.

Rakova, B., & Dobbe, R. (2023). Algorithms as social-ecological-technological systems: An environmental justice lens on algorithmic audits. *arXiv*. https://arxiv.org/abs/2305.05733

Reamer, F. G. (2023). Artificial intelligence in social work: Emerging ethical issues. *International Journal of Social Work Values and Ethics*, *20*(2), 52–71. https://doi.org/10.55521/10-020-205

Reamer, F. G. (2013). Social work in a digital age: Ethical and risk management challenges. *Social Work*, *58*(2), 163–172. https://doi.org/10.1093/sw/swt003

Sabry, F., Eltaras, T., Labda, W., Alzoubi, K., & Malluhi, Q. (2022). Machine learning for healthcare wearable devices: The big picture. *Journal of Healthcare Engineering*, *2022,* 4653923. https://doi.org/10.1155/2022/4653923

Segal, M. (2024). Confronting and managing ethical dilemmas in social work using ChatGPT. *European Journal of Social Work*, *28*(1), 155–167. https://doi.org/10.1080/13691457.2024.2377786

Silberberg, P. (2023). Animal ethics, animal welfare, and speciesism: Considerations for social work. *International Journal of Social Work Values and Ethics*, *20*(2), 72–111. https://doi.org/10.55521/10-020-206

Tronto, J. C. (2015). *Who cares? How to reshape a democratic politics*. Cornell University Press.

Tronto, J. C. (2013). *Caring democracy: Markets, equality and justice*. New York University Press.

United Nations Environment Programme (UNEP). (2024). *Measuring the environmental footprint of artificial intelligence*. UNEP.

WHO. (2024). The role of digital health technologies in women's health, empowerment and gender equality. https://iris.who.int/bitstream/handle/10665/376179/WHO-EURO2024-9293-49065-73153-eng.pdf?sequence=2

Zewe, A. (2025, January 17). Explained: Generative AI's environmental impact. *MIT News*. https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117

Zimmerman, M. E. (1995). The threat of ecofascism. *Social Theory and Practice*, *21*(2), 157–177. https://doi.org/10.5840/soctheorpract199521214

12 Bridging Worlds

Artificial Intelligence, ChatGPT, and Social-Work Ethics

Alberto Ballestero, Juan-Jesús Viscarret, and María-Jesús Úriz

Introduction

The rapid expansion of Artificial Intelligence (AI) technologies is reshaping nearly every professional domain, and social work is no exception. Among today's most widely used generative tools, ChatGPT—a large language model (LLM) developed by OpenAI—has drawn particular attention for producing human-like, context-aware text across diverse tasks. In social work, early adopters already use LLMs to streamline client documentation, generate plain-language resources, model therapeutic dialogues, and support policy analysis (James et al., 2024). These developments show real potential to extend service reach, lessen administrative burdens, and enhance data-driven decision-making.

Yet integrating AI into a profession rooted in relational practice and social justice raises ethical dilemmas requiring close scrutiny (Ahn et al., 2025). Other disciplines have advanced more rapidly: the American Medical Association (AMA, 2024) recommends physician oversight of diagnostic AI, and the American Bar Association (ABA, 2024) adopted Resolution 604 on algorithmic accountability. In contrast, social work's contribution to public and policy debates remains limited, even though it serves populations most vulnerable to technological harm.

Social workers now employ AI in clinical, administrative, educational, and advocacy contexts (Victor & Herrenkohl, 2025)—from mental-health chatbots and automated case notes to predictive analytics identifying individuals at risk of maltreatment (Mishna et al., 2021; Reamer, 2019; Saxena & Guha, 2023). Such tools, however, heighten concerns about informed consent, data protection, algorithmic bias, and erosion of professional judgment—especially in marginalized settings. Existing guidance, such as the NASW Standards for Technology in Social Work Practice (NASW et al., 2017), provides an initial framework but lacks specificity for generative models involving real-time text generation, probabilistic reasoning, and large-scale data ingestion.

Moreover, early studies show that ChatGPT's responses can be overly general for the complex ethical dilemmas social workers face. When presented with a case of suspected child neglect, for instance, ChatGPT offered generic best-practice steps but omitted jurisdiction-specific reporting laws and cultural context—both vital for ethical action (Segal, 2025b). These limitations underscore the need for

DOI: 10.4324/9781003627920-12
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

AI-specific ethical frameworks, curricular materials, and decision-making protocols adapted to social work (Reamer, 2023).

Accordingly, this chapter examines ChatGPT's ethical implications for social work. It reviews current applications, outlines five interlocking ethical challenges, assesses existing standards, and proposes a four-pillar strategy for responsible adoption. The argument is that the profession must move from reactive use to proactive governance—engaging practitioners, educators, developers, policymakers, and service users—to ensure that AI innovation reinforces, rather than undermines, social work's core commitments to human dignity, social justice, and the centrality of human relationships.

Current Applications of AI in Social Work

Social workers are increasingly integrating AI into diverse areas of practice, from micro-level clinical work to macro-level research and advocacy. Tools such as ChatGPT and other large-language models assist with clinical documentation, case-note summarization, decision support, educational material generation, and simulation of ethical dilemmas for teaching (Segal, 2025a). AI also contributes to grant writing, advocacy content development, and curriculum design in social-work education (Jarke & Breiter, 2019).

As Kumar and Singh (2023) note, tools like ChatGPT are being used in real-time clinical contexts to provide counselling support, streamline administrative records, and analyze client communications for emotional tone or mental-health indicators. These applications enhance efficiency, improve resource allocation, and facilitate early identification of client distress and tailored interventions.

At a more advanced stage, AI is also employed to automate risk assessments, deliver real-time crisis support, identify systemic biases in service provision, and predict professional burnout among social-service workers—though such tasks often rely on simpler algorithmic or decision-tree approaches rather than full machine learning (Grządzielewska, 2021).

These systems are increasingly integrated into child protection services and mental-health programs to assist in prioritizing interventions and resource allocation. Recent ethnographic research shows that algorithmic systems used in child welfare can cause harm when implemented without critical oversight, specifically by reinforcing existing institutional biases, misclassifying families as high risk, and generating increased surveillance or intrusive interventions for already marginalized groups. Such outcomes place additional pressure on everyday professional practice and can undermine trust between families and service providers (Gillingham, 2019; Saxena & Guha, 2023).

The use of AI in therapeutic settings is increasingly prominent. Mental-health apps such as Woebot and Wysa deliver automated, structured therapeutic dialogues based on evidence-based approaches like cognitive behavioural therapy (CBT). Woebot has demonstrated clinical efficacy in reducing depression and anxiety among young adults (Fitzpatrick et al., 2017), while Wysa enhances well-being through anonymous and scalable interaction (Inkster et al., 2018). Other tools,

including Virtual Hope Box and PTSD Coach, specifically assist military personnel and veterans with post-traumatic stress disorder (PTSD), anger management, and emotional regulation (Luxton et al., 2015). A typical application involves users accessing these tools through mobile apps or web platforms to complete self-guided mood tracking, grounding exercises, or CBT-based prompts; use is generally unsupervised, although some programs integrate optional clinician monitoring or referral pathways. Collectively, these innovations expand access to mental-health support, streamline service delivery, and offer practical resources for both clients and practitioners. However, their growing use also demands rigorous ethical scrutiny, given their influence on professional judgment and client outcomes.

Core Ethical Challenges

Informed Consent and Client Autonomy

Social workers have an ethical obligation to secure informed consent for interventions involving AI. In practice, AI enters social-work workflows through three modalities:

1 Back-office decision support (e.g., de-identified triage dashboards);
2 Augmented documentation (e.g., speech-to-text transcriptions summarized by ChatGPT);
3 Client-facing interaction (e.g., mental-health chatbots).

Consent requirements vary accordingly. Explicit, written consent is required whenever identifiable data are shared externally or when clients directly engage with AI agents (modalities 2 and 3) (Reamer, 2023). For internal, de-identified support (modality 1), a notice-and-opt-out policy is generally acceptable if human oversight and transparent AI data-use statements are maintained (Morley et al., 2020b).

To ensure genuine understanding, practitioners should provide layered explanations detailing benefits (reduced paperwork, quicker referrals), risks (misdiagnosis, hallucinations), data flows (storage, duration, ownership), and levels of human review (Kumar & Singh, 2023). This aligns with the NASW Code of Ethics (2021), emphasizing voluntary participation and preservation of professional judgment. For clients in vulnerable or mandated contexts, consent must transcend procedural compliance to uphold autonomy and should be paired with heightened professional oversight or supervision to ensure that AI-supported interventions do not compromise client rights or decision-making capacity.

Privacy and Confidentiality

Privacy and confidentiality remain central to ethical social work. AI systems raise significant data security concerns since many platforms store, analyze, or reuse data for model training (Ahn et al., 2025). ChatGPT, for instance, can retain interaction data unless strict safeguards are implemented. Compliance with data-protection

laws is essential; social workers must inform clients, obtain consent, and ensure transparency about how personal data are handled (Badillo-Díaz, 2024). Such practices uphold trust and align with social work's duty to protect client information.

Frameworks like the General Data Protection Regulation (GDPR) impose stringent obligations on data processing (Kosseff, 2019). Mechanisms such as Data Protection Impact Assessments (DPIAs), robust encryption, anonymization, vendor contracts, and external audits help identify and mitigate privacy risks. Transparent client communication and documentation of AI data workflows further reinforce accountability. These combined organizational and technical safeguards ensure that privacy and confidentiality principles are operationalized in daily practice (AEPD, 2021).

Algorithmic Bias and Fairness

AI systems often mirror and reproduce biases embedded in historical data, potentially exacerbating inequalities. Wach et al. (2023) note that, in the absence of robust market regulation, large-scale systems such as ChatGPT can amplify stereotypes and generate discriminatory outcomes, especially when embedded in automated social-care services. Patton et al. (2023) add that large-language models are trained on corpora, which are extensive collections of text data drawn from diverse sources such as books, websites, social media, and other written materials, that under—or misrepresent marginalized voices—people with disabilities, older adults, BIPOC communities, and the Global South—thereby reinforcing epistemic injustice unless actively countered by antiracist design. In child-welfare and criminal-justice contexts, biased datasets have caused algorithmic tools to flag Black and disabled families for investigation at disproportionately high rates (Eubanks, 2018; Saxena & Guha, 2023).

To uphold equity, social workers can implement a structured four-step protocol to mitigate bias in AI systems used in practice:

1 Representational Audit Before Deployment. Examine how the training corpus reflects the demographics of the target population and document dataset provenance, coverage, and limitations using datasheets. Address gaps by incorporating community-curated or counterfactually augmented data (CAD) to strengthen robustness against spurious associations, embedding this process within an end-to-end internal audit (Gebru et al., 2021; Kaushik et al., 2020; Wachter et al., 2018; Raji et al., 2020).
2 Fairness Metrics as Gatekeeping Tests. Apply quantitative parity tests—equality of opportunity (true-positive rate parity) and equalized odds (joint parity of TPR and FPR)—disaggregated by race, disability, and age. Deviations exceeding ±5 percentage points (a practical, not universal, threshold) should trigger model retraining and expert review. This aligns with the technical literature and internal auditing frameworks (Hardt et al., 2016; Raji et al., 2020).
3 Lived-Experience Review Board. Create an advisory panel of service users and affected communities with authority to veto deployment when audits reveal

disproportionate risks. Such participatory approaches, supported by data-justice and algorithmic-impact frameworks, ensure that lived expertise informs oversight (D'Ignazio & Klein, 2020; Reisman et al., 2018).

4 Post-Deployment Monitoring Ledger. Record each high-stakes AI recommendation, enabling quarterly analyses of disparate impact and model drift. Findings should appear in public model cards and be shared with regulators, consistent with the NIST AI RMF 1.0 and the EU AI Act on post-market monitoring for high-risk systems (Mitchell et al., 2019; National Institute of Standards and Technology [NIST], 2023; European Union, 2024).

Adopting this protocol operationalizes social work's anti-oppressive mandate, reframing evaluation from "Is the tool accurate?" to "For whom and at whose cost?" (James et al., 2024). In practice, such a protocol is typically applied within high-stakes public-service settings—for example, child-welfare agencies evaluating a risk-assessment model before deployment—and is usually funded through a combination of agency operating budgets, state or municipal innovation funds, or externally supported research–practice partnerships. When parity fails and redress is infeasible, ethical use requires halting or discontinuing the AI system in favour of human-led alternatives.

Transparency and Accountability

AI systems in social services—such as LLM-based case summarizers, triage dashboards, and risk scorers—often operate as black boxes, due to technical opacity or proprietary restrictions (Pasquale, 2015). Existing legal frameworks still apply; agencies and professionals remain accountable under duty-of-care, negligence, administrative law, and data-protection statutes such as the EU's General Data Protection Regulation (GDPR) and the U.S. Health Insurance Portability and Accountability Act (HIPAA). Opacity complicates compliance, hindering explanations to clients, justification of decisions, and detection of errors. This creates an accountability gap between designers and frontline practitioners, who feel answerable to clients without understanding system reasoning.

While AI-specific regulation remains limited, accountability in public-service contexts already rests on professional and legal duties. However, LLM-based tools (e.g., ChatGPT-derived systems) can weaken these safeguards when their internal logic remains invisible. To strengthen accountability, agencies should require pre-deployment assessments (Data Protection Impact Assessments [DPIAs], algorithmic-impact reviews), interpretability artifacts (datasheets, model cards), immutable logs of inputs, outputs, and human overrides, and vendor contracts ensuring audit rights and prompt error reporting. These mechanisms collectively preserve explainability and traceability in AI-supported decision-making.

Automation is not an on/off switch but a sliding scale. Drawing on levels-of-automation taxonomies, we propose a ceiling for high-stakes social-work contexts: "AI recommends, human decides". In practice, this means (a) no automated adverse actions without documented human review; (b) mandatory human sign-off

for high-risk recommendations; and (c) second-review checkpoints when model scores exceed predefined thresholds. For example, in child protection, an external audit of a screening tool identified disparate impacts across racial groups; the agency responded by reweighting features, adding human-override checkpoints, and documenting decision rationales alongside model scores, thereby retaining final authority while improving explainability and auditability (Chouldechova et al., 2018). However, there is a growing body of evidence that humans tend to increasingly 'trust' AI recommendations, particularly when under time pressure (Yang et al., 2025).

Although many studies report "positive" evaluations of AI, reviews consistently find major omissions regarding fairness, explainability, and accountability. The CONSORT-AI and SPIRIT-AI extensions were designed to fill these gaps by standardizing reporting on clinical integration, code availability, error analysis, and subgroup performance (Liu et al., 2020; Cruz Rivera et al., 2020). Yet systematic reviews reveal weak adherence to these frameworks, high risk of bias, and limited subgroup analysis. For instance, Lu et al. (2022) found only 39% adherence to reporting items among models from a single vendor, while Plana et al. (2022) documented wide variability and underrepresentation of vulnerable groups.

Accordingly, AI deployments should be conditioned on predefined thresholds for calibration, error rates, and disparity, with pause/rollback mechanisms when these are breached. Continuous post-deployment monitoring should align with the NIST AI RMF (2023) and the EU AI Act (2024), which classifies systems by four risk levels: minimal, limited (transparency obligations), high, and unacceptable.

Professional Competence and Human Judgment

Ethical AI use in social work—for example, employing automated transcription tools to streamline documentation or using decision-support dashboards to supplement, rather than replace, practitioner judgment—depends on professional competence and institutional safeguards. Practitioners must understand AI's capacities and limits, using outputs as inputs to—not substitutes for—professional judgment. Overreliance risks weakening empathy and critical thinking (Badillo-Díaz, 2024). Training remains essential, but structural pressure—cost-cutting, performance targets, and risk aversion—often drives substitution rather than augmentation.

Organizations should codify protective measures: apply a "AI recommends, human decides" rule for high-stakes decisions; forbid automated adverse actions without documented review; require human sign-off for high-risk outputs; and ensure staffing ratios that preserve reflective practice. Oversight logs should track overrides and rationales.

The ethical boundary between augmentation and replacement is crossed when human review becomes perfunctory, eligibility decisions are automated, or metrics reward alignment with model outputs over client-centred reasoning. Such conditions erode professional judgment and relational ethics (Segal, 2025b). Sustaining competence therefore requires continuous education, interdisciplinary consultation, ethics rounds, and governance incentives ensuring AI consistently augments rather than replaces human expertise.

Specific Concerns with ChatGPT

Superficial Analysis

Social workers report that ChatGPT's responses to complex ethical or contextual issues often lack depth and reflexivity. While it can outline key principles or thematic summaries, its probabilistic design favours generalization over contextual understanding (Patton et al., 2023). Consequently, ChatGPT struggles to address the nuanced interplay of personal history, culture, and institutional context central to ethical practice (Ahn et al., 2025; Reamer, 2023). Such superficiality may weaken professional judgment and client trust.

Limited Emotional Understanding

Current large-language models—including ChatGPT—cannot perceive tone or non-verbal cues; simulated empathy arises from lexical patterns, not emotional awareness. An adjacent emotion-AI field seeks to infer affect from text, voice, or facial micro-expressions (McStay, 2018), but cross-cultural studies reveal high error rates and limited validity (Barrett et al., 2019). The EU Artificial Intelligence Act (Regulation EU 2024/1689) now prohibits most workplace deployments of emotion-recognition systems except where essential for safety. Peer-reviewed evidence that would support parity with human clinical judgment has not been forthcoming to date. Hence, in therapeutic contexts where affective attunement is critical, ChatGPT should serve only as an adjunct—for psychoeducation or normalization exercises—while clinicians retain full responsibility for interpreting emotion and tailoring care (Fitzpatrick et al., 2017; Inkster et al., 2018; Ahn et al., 2025).

Legal and Regulatory Considerations

Ethical principles in social work—autonomy, beneficence, non-maleficence, and justice—are articulated in the NASW Code of Ethics (2021) and partly codified in law (Banks, 2022). However, ChatGPT does not reliably recognize jurisdiction-specific statutes, risking incomplete or misleading guidance (Segal, 2025b). For example, when asked about mandatory reporting in Ontario, it failed to mention the Child, Youth and Family Services Act, whose omission could expose practitioners to liability. Human ethics committees, by contrast, integrate both ethical and statutory reasoning (Reamer, 2023). Practitioners should thus treat AI outputs as preliminary and cross-check them against legal frameworks or use retrieval-augmented systems fine-tuned for local statutes.

Professional Standards and Guidelines

NASW Technology Standards

The NASW, ASWB, CSWE, and CSWA Standards for Technology in Social Work Practice (2017) remain the profession's principal ethical reference on

digital engagement. They address confidentiality, consent, boundaries, accessibility, supervision, and competence across four domains: (1) public information (agency websites, social media); (2) technology-based services (telepractice, emergency planning, plain-language consent); (3) data management (privacy, cross-jurisdictional storage); and (4) administration and supervision (training, vendor management, record-keeping).

While foundational, these standards predate generative AI and thus omit key elements: (a) acceptable uses of LLMs for real-time text generation or summarization; (b) fairness, bias, and calibration requirements; (c) expectations for data provenance, transparency artifacts (datasheets, model cards), and post-deployment monitoring; (d) vendor accountability for logs, audits, and incident reporting; and (c) authority limits consistent with "AI recommends, human decides".

Need for AI-Specific Guidelines

Current technology standards fall into two layers: technical norms (e.g., ISO/IEC 42001; IEEE ECPAIS certification) and professional ethics codes (NASW et al., 2017). While the former address cybersecurity and risk management, and the latter affirm confidentiality and consent, neither was built for large-language-model ecosystems.

The EU Artificial Intelligence Act (Regulation EU 2024/1689) adds a legal layer, classifying most social-service AI as high risk and mandating risk assessments and human oversight. Yet enforcement mechanisms and sector-specific audit metrics remain underdeveloped. As James et al. (2024) observe, the profession lags not from lack of guidance but from insufficient technical specificity for frontline application.

Scholars thus call for AI-specific ethical toolkits translating broad principles—beneficence, autonomy, justice—into operational mechanisms via bias-audit thresholds, prompt-engineering safeguards, and participatory oversight protocols (Morley et al., 2020b; Reamer, 2023). Without them, institutions risk ethics blue-washing: symbolic compliance without protection for clients or staff. The gap is evident in education. Shklarski and Ray (2024) note that universities lack coherent policies on ChatGPT use, leaving uncertainty about integrity and equitable access. Coordinated policies—co-developed by regulators, professional bodies, and engineers—are vital to embed transparency, accountability, and equity at the system level, not merely in codes of conduct.

Strategies for Ethical AI Implementation

To address the ethical challenges of AI in social work, practitioners, organizations, and educational institutions must adopt deliberate strategies aligning innovation with professional values. Four interrelated action areas can guide this integration: ethical frameworks, training and education, human–AI collaboration, and continuous monitoring and evaluation.

Establishing Ethical Frameworks

Professional associations (e.g., NASW, 2017; AASW, 2020), major service agencies, and universities share responsibility for developing context-sensitive ethical frameworks grounded in respect for human dignity, social justice, and the primacy of human relationships. These frameworks should specify protocols for informed consent, data governance, and the limits of AI-supported intervention.

A practical model is the digital-ethics committee, a multidisciplinary body reviewing deployments, evaluating social impact, and ensuring consistency with institutional policy and professional standards (Floridi & Cowls, 2019). Morley et al. (2020a) propose linking concrete tools—bias audits, value-sensitive design, and data-nutrition labels—to each stage of the AI lifecycle, from data collection to post-deployment monitoring. Such participatory and iterative mechanisms align with social work's emphasis on procedural justice and stakeholder inclusion. In high-stakes areas such as child welfare or mental health, regular algorithmic audits remain essential for bias detection and correction (Saxena & Guha, 2023).

Training and Education

Ethical AI integration requires dual-track capacity-building: technical competence (understanding how systems ingest and process data) and ethical literacy (knowing when and with what safeguards to apply them) (Morley et al., 2020a). Studies show a widening competence gap: many practitioners lack formal training in AI or digital ethics (Ahn et al., 2025), and few can critically assess ChatGPT's recommendations (Goldkind et al., 2023). This gap is consequential, because without the ability to interrogate AI outputs, practitioners cannot reliably identify errors, biases, or inappropriate recommendations, weakening the ethical foundations on which responsible AI use depends.

Professional associations and universities must therefore embed accessible learning pathways—micro-credentials, practice labs, and mandatory CPD—into professional development. In higher education, Shklarski and Ray (2024) find educators appreciate ChatGPT's pedagogical potential but fear "false mastery". Program directors should (a) include syllabus statements defining acceptable AI use, (b) offer faculty training in prompt engineering and bias awareness, and (c) require reflective assignments linking AI outputs to social justice analysis. These measures sustain ethical reflection rather than uncritical adoption.

Human–AI Collaboration

AI should augment, not replace, professional judgment. A human–AI collaboration model preserves relational and reflective practice while leveraging computational capacities (James et al., 2024). Relevant organizations include child welfare, income-support, hospital and behavioural-health units, schools, courts, nonprofits, and vendors supplying AI tools.

Goldkind et al. (2023) recommend determining which tasks may be ethically automated (data entry, transcription, summarization) and which require human

oversight (eligibility decisions, service planning, crisis response, adverse actions). Policies should clearly define approval thresholds, second-review procedures for high-stakes cases, documentation duties, and client opt-out options. Cross-disciplinary ethics committees ensure AI deployment aligns with the mission of social work rather than efficiency targets (Reamer, 2023; James et al., 2024). Social workers must retain final authority in all decisions, using AI outputs as supplementary information subject to human verification.

Continuous Monitoring and Evaluation

Ethical AI implementation depends on ongoing monitoring of performance, outcomes, and unintended effects. Organizations should maintain detailed usage logs, collect feedback from staff and clients, and assess whether tools achieve their goals without compromising ethics. Transparency and interdisciplinary review foster accountability by defining responsibilities, enabling audits, and clarifying remediation pathways. Iterative evaluation ensures systems adapt as technologies, regulations, and social contexts evolve (Floridi et al., 2018; NIST, 2023; Mitchell et al., 2019).

Responsibilities are distributed across multiple layers:

a Operational leaders manage daily oversight and corrective actions.
b Quality and compliance teams (including data-protection officers and ethics committees) conduct audits and ensure regulatory alignment.
c Frontline workers and supervisors provide feedback and retain override authority.
d Community and lived-experience panels assess impacts and can request suspension or redesign.
e Independent reviewers validate methods through external audits or academic partnerships.

Operational mechanisms should include immutable logs linking model versions, inputs, outputs, and human overrides; fairness and safety dashboards tracking error rates, calibration, and disparities; drift detection alerts; and accessible client feedback channels. Structured evaluations—pre-registered plans, holdout testing, controlled shadow deployments, and debriefs—should inform governance gates for retraining or rollback. Transparent reporting through public model cards and periodic summaries builds community trust (Mitchell et al., 2019; NIST, 2023; EU, 2024).

An illustrative cycle includes real-time logging, weekly automated checks, monthly bias reviews, quarterly interdisciplinary board meetings, and annual external audits. These practices make oversight continuous, participatory, and verifiable (Floridi et al., 2018; Reisman et al., 2018).

Future Directions

AI's integration into social work is still formative, offering innovation but also risk. The field faces a pivotal question: how to harness these tools responsibly while

preserving the ethical foundations and human relationships central to practice (Ahn et al., 2025; Reamer, 2023). Core principles—dignity, justice, and well-being—must remain central as new AI-specific ethical guidelines evolve, addressing algorithmic opacity, machine judgment limits, and protection of vulnerable populations (James et al., 2024).

These guidelines should combine empirical research and participatory dialogue, incorporating perspectives from clients, practitioners, educators, and affected communities. Continued investment in digital literacy and professional development is vital so that social workers can assess AI's social implications using ethical reasoning (Morley et al., 2020a). Universities must integrate AI ethics into curricula to prepare future professionals.

The profession also needs robust accountability mechanisms and research on AI's effects on decision-making, equity, and outcomes. Interdisciplinary collaboration—with technologists, ethicists, and communities—will help ensure AI serves the public good rather than reproducing structural injustice (Floridi & Cowls, 2019; Eubanks, 2018).

Ultimately, the question is not whether social work should use AI, but how—ethically, critically, and collaboratively. With safeguards, AI can empower practitioners and make services more inclusive; without them, it risks deepening marginalization. The future depends on social work's ability to lead this transformation, embedding oversight through staged, risk-sensitive pathways:

a Register all AI deployments and associated use cases and conduct participatory impact assessments to clarify purpose, scope, and affected groups.
b Require transparency artifacts (model cards, procurement checklists).
c Run shadow trials with harm safeguards and go/no-go fairness gates.
d Maintain continuous monitoring, drift reporting, and public transparency.

Responsibilities are distributed among:

a Professional associations setting field guidance;
b Regulators and licensing bodies enforcing standards;
c Agency leadership resourcing governance;
d Quality, compliance, and data teams auditing dashboards;
e Frontline practitioners providing feedback and override authority;
f Community advisory boards reviewing impacts;
g Independent auditors validating methods;
h Universities and training institutions building capacity through ethics education and continuing development.

Conclusion

The integration of AI—particularly tools like ChatGPT—marks a pivotal juncture for social work. While AI offers new possibilities in service delivery, documentation, education, and mental-health support, its adoption must remain anchored in critical ethical reflection and the profession's core values.

AI systems such as ChatGPT, predictive-risk models, and sentiment analysers are already embedded in social services. Social work's influence therefore extends from upstream processes—data selection and system design—to downstream oversight through contract negotiation, audit triggers, and public reporting. By engaging at these leverage points, social workers can ensure that the five ethical priorities outlined in this chapter—informed consent, confidentiality, bias mitigation, accountability, and competence—are operationalized rather than merely aspirational (Reamer, 2023).

If professional bodies, educators, and frontline agencies co-lead this agenda, AI can expand reach while preserving relational integrity. Without such leadership, algorithmic systems may follow commercial logics that deepen inequities and erode professional judgment. The real question is not whether social work will engage with AI, but whether it will shape or be shaped by it.

This chapter has identified five central ethical concerns—informed consent, confidentiality, bias, accountability, and professional competence—and examined ChatGPT's limitations in emotional intelligence, contextual sensitivity, and legal awareness. It has proposed four strategies for ethical implementation: creating institution-specific frameworks, enhancing digital education, fostering human–AI collaboration, and establishing continuous evaluation systems. These must be reinforced by updated professional standards and sustained interdisciplinary dialogue among practitioners, educators, developers, and service users.

Ultimately, the goal is not to adapt social work passively to technological change but to shape AI's role in alignment with human dignity, social justice, and ethical responsibility. Guided by these principles, AI can strengthen the mission of social work; without them, it risks amplifying inequality and undermining the very relationships that define ethical and effective practice.

References

Agencia Española de Protección de Datos. (AEPD) (2021). *Audit requirements for personal data processing activities involving AI*. https://www.aepd.es/guides/audit-requirements-for-personal-data-processing-activities-involving-ai.pdf

Ahn, E., Choi, M., Fowler, P., & Song, I. H. (2025). Artificial intelligence (AI) literacy for social work: Implications for core competencies. *Journal of the Society for Social Work and Research*, *16*(1). https://doi.org/10.1086/735187

American Bar Association, Standing Committee on Ethics and Professional Responsibility. (2024, July 29). *Formal Opinion 512: Generative artificial intelligence tools*. American Bar Association. https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf

American Medical Association. (2024). *Augmented intelligence development, deployment, and use in health care* (House of Delegates policy adopted November 12, 2024). American Medical Association. https://www.ama-assn.org/system/files/ama-ai-principles.pdf

Australian Association of Social Workers. (2020). *Code of ethics*. AASW.

Badillo-Díaz, M. (2024, October 18). *The ethical implications of AI in social work*. The AI Social Worker. https://www.theaisocialworker.com/blog/the-ethical-implications-of-ai-in-social-work

Banks, S. (2022). *Ethics and values in social work* (5th ed.). Bloomsbury Academic.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest, 20*(1), 1–68. https://doi.org/10.1177/1529100619832930

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proceedings of Machine Learning Research, 81*, 134–148. https://proceedings.mlr.press/v81/chouldechova18a/chouldechova18a.pdf

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., Calvert, M. J., & The SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nature Medicine, 26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. MIT Press.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behaviuor therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomised controlled trial. *JMIR Mental Health, 4*(2), e19. https://doi.org/10.2196/mental.7785

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review, 1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Gillingham, P. (2019). Can predictive algorithms assist decision making in social work with children and families? *Child Abuse Review*, *28*(2), 114–126. https://doi.org/10.1002/car.2547

Goldkind, L., Wolf, L., Glennon, A., Rios, J., & Nissen, L. (2023). The end of the world as we know it? ChatGPT and social work. *Social Work, 69*(1), 103–105. https://doi.org/10.1093/sw/swad044

Grządzielewska, M. (2021). Using machine learning in burnout prediction: A survey. *Child and Adolescent Social Work Journal, 38*, 175–180. https://doi.org/10.1007/s10560-020-00733-w

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems 29 (NeurIPS 2016)*. https://arxiv.org/abs/1610.02413

IEEE. (2024). *ECPAIS certification criteria: Ethics Certification Programme for Autonomous and Intelligent Systems.* IEEE Standards Association.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. *JMIR mHealth and uHealth, 6*(11), e12106. https://doi.org/10.2196/12106

International Organisation for Standardisation. (2019). *ISO/IEC 27701:2019—Security techniques—Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management—Requirements and guidelines*. ISO. https://www.iso.org/standard/71670.html

ISO/ IEC. (2023). *ISO/IEC 42001:2023 Artificial intelligence—Management system.* International Organisation for Standardisation & International Electrotechnical Commission.

James, P., Lal, J., Liao, A., Magee, L., & Soldatic, K. (2024). Algorithmic decision-making in social work practice and pedagogy: Confronting the competency/critique dilemma. *Social Work Education*, *43*(6), 1552–1569. https://doi.org/10.1080/02615479.2023.2195425

Jarke, J., & Breiter, A. (2019). The datafication of education. *Learning, Media and Technology, 44*(1), 1–3. https://doi.org/10.1080/17439884.2019.1573833

Kaushik, D., Hovy, E., & Lipton, Z. C. (2020). Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations (ICLR 2020)*. https://openreview.net/pdf?id=Sklgs0NFvr

Kosseff, J. (2019). *Cybersecurity law*. Wiley.

Kumar, A., & Singh, N. K. (2023). Use of artificial intelligence in social work practice. *Indian Journal of Health Social Work, 5*(2), 2–8.

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., Denniston, A. K., & The SPIRIT- AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. *Nature Medicine, 26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Lu, J. H., Callahan, A., Patel, B. S., Morse, K. E., Dash, D., Pfeffer, M. A., & Shah, N. H. (2022). Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: A systematic review. *JAMA Network Open, 5*(8), e2227779. https://doi.org/10.1001/jamanetworkopen.2022.27779

Luxton, D. D., McCann, R. A., Bush, N. E., Mishkind, M. C., & Reger, G. M. (2015). mHealth for mental health: Integrating smartphone technology in behaviuoral healthcare. *Professional Psychology: Research and Practice, 42*(6), 505–512. https://doi.org/10.1037/a0024485

McStay, A. (2018). *Emotional AI: The rise of empathic media*. Sage.

Mishna, F., Sanders, J. E., Sewell, K. M., & Milne, E. (2021). Teaching note—Preparing social workers for the digital future of social work practice. *Journal of Social Work Education, 57*(sup1), s19–s28. https://doi.org/10.1080/10437797.2021.1912676

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* '19)*, Atlanta Georgia (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020a). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020b). The ethics of AI in health care: A mapping review. *Social Science & Medicine, 260*, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

National Association of Social Workers [NASW], Association of Social Work Boards, Council on Social Work Education, & Clinical Social Work Association. (2017). *NASW, ASWB, CSWE, & CSWA standards for technology in social work practice*. https://www.socialworkers.org/Practice/NASW-Practice-Standards-Guidelines

National Association of Social Workers. (2017). *Standards for technology in social work practice*. NASW.

National Association of Social Workers. (2021). *National association of social workers code of ethics*. https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English

National Institute of Standards and Technology (NIST) (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Patton, D. U., Landau, A. Y., & Mathiyazhagan, S. (2023). ChatGPT for social work science: Ethical challenges and opportunities. *Journal of the Society for Social Work and Research, 14*(3), 553–561. https://doi.org/10.1086/726042

Plana, D., Shung, D. L., Grimshaw, A. A., Saraf, A., Sung, J. J. Y., & Kann, B. H. (2022). Randomised clinical trials of machine learning interventions in health care: A systematic review. *JAMA Network Open, 5*(9), e2233946. https://doi.org/10.1001/jamanetworkopen.2022.33946

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*, Barcelona, Spain. Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Reamer, F. G. (2019). Social work education in a digital world: Technology standards for education and practice. *Journal of Social Work Education, 55*(3), 420–432. https://doi.org/10.1080/10437797.2019.1567412

Reamer, F. G. (2023). Artificial intelligence in social work: Emerging ethical issues. *International Journal of Social Work Values and Ethics*, *20*(2), 52–71. https://doi.org/10.55521/10-020-205

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic impact assessments: A practical framework for public agency accountability*. AI Now Institute. https://ainowinstitute.org/wp-content/uploads/2023/04/aiareport2018.pdf

Saxena, D., & Guha, S. (2023). Algorithmic harms in child welfare: Uncertainties in practice, organisation, and street-level decision-making [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2308.05224

Segal, M. (2025a). Confronting and managing ethical dilemmas in social work using ChatGPT. *European Journal of Social Work, 28*(1), 155–167. https://doi.org/10.1080/13691457.2024.2377786

Segal, M. (2025b). Social workers' evaluation of ChatGPT for solving ethical dilemmas within the limits of confidentiality. *Journal of Social Work Practice*. Advance online publication. https://doi.org/10.1080/02650533.2025.2480092

Shklarski, L., & Ray, K. (2024). Navigating the dilemma: Use of ChatGPT in social work education. *Journal of Teaching in Social Work, 44*(5), 476–494. https://doi.org/10.1080/08841233.2024.2408299

Victor, B. G., & Herrenkohl, T. I. (2025). Introducing a new *JSSWR* specialty series on artificial intelligence and social work. *Journal of the Society for Social Work and Research, 16*(1), 5–8. https://doi.org/10.1086/735286

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review, 11*(2), 7–30. https://doi.org/10.15678/EBER.2023.110201

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology,31*(2),841–887.https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf

Yang, R., Li, S., Qi, Y., Liu, J., He, Q., & Zhao, H. (2025). Unveiling users' algorithm trust: The role of task objectivity, time pressure, and cognitive load. *Computers in Human Behaviuor Reports*, *18*. https://doi.org/10.1016/j.chbr.2025.100667

13 A Cruel Kindness? The Limitations of Logic-Based Artificial Intelligence in Navigating Ethical Dilemmas within the Human Services Context

Clement Chihota and Tariro Chihota

Introduction

Artificial intelligence (AI) has been defined by the American National Standard Dictionary of Artificial Intelligence (cited in Atherton et al., 2023) as "… the capability of a device to perform functions that are normally associated with human intelligence such as reasoning, learning, and self-improvement". Similarly, Russell and Norvig (2016) describe AI as the design and construction of intelligent agents that perceive their environments and act upon them. Across healthcare, social and community services, and counselling, AI systems are increasingly embedded in professional decision-making processes, particularly in areas such as risk assessment, triage, and resource allocation.

Despite these developments, the capacity – or limitation – of AI to navigate ethically complex situations remains a pressing concern. Practitioners in human services routinely contend with moral ambiguity, balance competing obligations, and respond empathetically to the emotional and relational realities of individuals experiencing vulnerability (Donaher, 2021). While AI excels in pattern recognition, large-scale data processing, and logical consistency, it lacks emotional intelligence, moral intuition, and sensitivity to thick contextual and relational detail (Buolamwini & Gebru, 2018; Inayatullah, 2001). These limitations become most consequential in contexts where ethical judgement and human relationships are central, and where the consequences of decision-making are deeply personal and enduring.

Defining AI in This Chapter

AI is not monolithic. This chapter focuses on systems currently deployed in human services that are logic- or data-driven, including rule-based decision-support tools, expert systems, and machine-learning models (including Large Language Models, LLMs) trained on large datasets to identify patterns, forecast outcomes, and generate text. Hypothetical artificial general intelligence (AGI) and superhuman AI fall outside the scope of this discussion. Concentrating on present and near-term technologies enables a more precise assessment of their ethical limits and practical implications for social care (Russell & Norvig, 2016; Floridi et al., 2018).

DOI: 10.4324/9781003627920-13
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

Ethics, Morality, and Norms

To avoid conceptual conflation, this chapter distinguishes among ethics, morality, and norms. Ethics refers to formalised frameworks and professional codes that articulate duties, principles, and standards of conduct (Floridi et al., 2018). Morality, by contrast, denotes lived, context-dependent judgement shaped by experience, empathy, and moral intuition—an orientation foregrounded in virtue ethics and in psychological accounts of moral reasoning (Gilligan, 1982; Haidt, 2001). Norms are socially enforced expectations that vary across communities, cultures, and institutional settings, and may or may not align with ethical principles or individual moral convictions (Tronto, 1993; Dancy, 2001).

The central question, therefore, is not whether AI systems can reproduce or cite ethical rules, but whether they can engage in moral judgement: an activity that requires sensitivity to context, relationships, and character. This perspective draws on virtue ethics, care ethics, and moral particularism, each of which resists the assumption that fixed rules alone can resolve complex human dilemmas (Gilligan, 1982; Tronto, 1993; Dancy, 2001).

Ethical Dilemmas in Human Services

Ethical dilemmas in human services are rarely clear-cut. Social workers, counsellors, and healthcare providers routinely confront conflicts of duty, uncertain outcomes, and emotionally charged contexts (Baillergeau & Duyvendak, 2016; Hillman & Chen, 2008). AI systems can be engineered to reflect teleological reasoning (maximising beneficial outcomes) and deontological reasoning (adherence to rules and duties), and in limited ways to gesture toward virtue-ethical considerations. However, when these frameworks are applied rigidly or in isolation, they remain insufficient. Human moral reasoning is heuristic, intuitive, and deeply contextual; it incorporates tacit knowledge, lived experience, and empathic attunement – qualities that logic-driven AI lacks or can only approximate superficially (Resnik, 2017; Haidt, 2001; Gigerenzer, 2008).

Consider a child protection case in which an AI risk model recommends removal based on aggregated indicators and historical patterns. Such a recommendation may be internally consistent and statistically defensible, yet blind to household strengths, the psychological harms of removal, or relational supports that could mitigate risk through less intrusive interventions. This exemplifies a "cruelly kind" outcome: a decision that is logic-compliant and procedurally justified, but ethically damaging in lived experience (O'Neil, 2016; Obermeyer et al., 2019).

Specialised Ethical AI

Some AI systems can generate structured ethical analyses – for example, producing arguments in utilitarian, deontological, or virtue-ethical terms, or flagging tensions among competing principles. Such tools can support reflective practice by clarifying trade-offs and surfacing implicit assumptions. However, generating ethical arguments is not equivalent to exercising moral judgement. In human services,

moral discernment requires sensitivity to emotion, relationships, and situational nuance, as well as the capacity to tolerate uncertainty – areas in which current AI systems remain fundamentally limited. Accordingly, specialised ethical AI should be framed as an aid to deliberation rather than a substitute for human judgement (Mittelstadt et al., 2016; Jobin, Ienca, & Vayena, 2019).

Capacity versus Normative Authority (and Why It Matters)

A critical distinction must be drawn between technical capacity and normative authority. The former concerns whether AI systems can generate, simulate, or evaluate ethical reasoning; the latter concerns whether they ought to be delegated moral agency or regarded as possessing moral standing. Even as AI systems increasingly support ethical analysis, it does not follow that they should be accorded normative authority. Scholarship on epistemic authority and accountability cautions against delegating ethically consequential decisions to entities that lack responsibility, liability, answerability, and the capacity for moral growth (Nemitz, 2018; Selbst et al., 2019). Human practitioners must therefore retain ultimate responsibility for morally significant decisions, ensuring that answerability remains with persons or accountable institutions rather than with models, metrics, or automated outputs alone (NASW, 2021; Floridi et al., 2018).

Human–AI Moral Symbiosis and the Normalisation of Assumptions

Moving beyond a binary "for or against" stance, it is evident that AI already shapes human moral agency. Decision-support systems influence how practitioners assess risk, justify interventions, and allocate scarce resources. Ethical assumptions encoded within models – such as what counts as harm, success, or fairness – can become normalised, subtly redistributing responsibility among individuals, institutions, and computational systems. In this sense, AI co-constitutes the moral environment without itself possessing moral agency. Recognising this symbiosis is essential to maintaining ethical accountability and preventing drift toward technocratic rationalities that marginalise plural values and lived experience (Mittelstadt et al., 2016; Veale & Brass, 2019).

Bias, Equity, and Contestability

AI systems can perpetuate or exacerbate inequities when trained on biased datasets or when optimisation relies on problematic proxies (for example, cost standing in for need). Empirical research documents disparities across domains, including facial recognition errors by skin tone and gender, biased risk scores in healthcare triage, and opacity that undermines contestability and informed consent (Buolamwini & Gebru, 2018; Obermeyer et al., 2019). In human services, such biases may lead to misclassification, disproportionate surveillance, and over-intervention in already marginalised communities. These failures underscore the importance of fairness, transparency, and robust governance (O'Neil, 2016; Jobin, Ienca, & Vayena, 2019). Across applied domains, from automated

driving to clinical decision support, technical improvement often coexists with unresolved ethical questions concerning responsibility, explainability, and trust – reinforcing the imperative that AI complement rather than displace human oversight (Nemitz, 2018; Floridi et al., 2018).

Moral Status and Agency

Current AI systems lack sentience, self-awareness, and intentionality; they neither experience harm nor possess interests in a morally relevant sense. As such, they do not qualify as moral patients or moral agents under prevailing ethical frameworks (Bryson, 2010; Shevlin, 2021). Extending moral status to AI would require profound cognitive, social, and emotional capacities that are wholly absent from existing architectures. Moreover, AI's apparent "morality" is ontological rather than axiological: systems reflect and stabilise values already present in human societies rather than originating independent value commitments. Without axiological agency – the capacity to justify why particular principles ought to hold – AI remains ethically derivative and prone to reproducing unjust norms prevalent in its training data.

Moral development theory further sharpens this concern. On Kohlberg's (1981) account, an autonomous moral agent must be capable of post-conventional reasoning, guided by universal ethical principles rather than conformity to prevailing norms. Historical examples of moral courage – such as Oskar Schindler's decision to protect Jewish workers during the Holocaust despite intense social and political pressure to comply with genocidal state policy – illustrate the capacity to transcend socially normalised injustice. By contrast, AI systems that replicate dominant patterns risk amplifying harmful templates when societies themselves normalise oppression. Absent an independent axiological basis and post-conventional moral reasoning, AI remains ethically subordinate and potentially complicit.

Implications for Practice

AI will continue to become faster, cheaper, and more capable across pattern discovery, forecasting, and language generation. Yet its architecture remains inherently derivative: such systems learn from human-produced data and objectives, generalising patterns rather than originating values. Consequently, contemporary AI cannot function as an independent moral agent capable of post-conventional reasoning or of justifying first-order ethical principles (Kohlberg, 1981; Floridi et al., 2018). This limitation establishes a clear division of labour in professional practice. AI should be understood as an epistemic instrument – highly effective at surfacing signals, revealing trade-offs, and supporting complex analysis – while normative authority over ethically consequential decisions must remain with morally accountable practitioners and organisations (Mittelstadt et al., 2016; Nemitz, 2018). Framed in this way, AI can strengthen ethical decision-making by augmenting, rather than replacing, human judgement, provided its deployment is accompanied by robust governance capable of preventing "cruelly kind" outcomes that are procedurally defensible yet harmful in lived experience (O'Neil, 2016; Obermeyer et al., 2019).

Within social work and the broader human services, AI's most valuable role lies in enhancing clinical reasoning, widening practitioners' field of view, and strengthening reflective capacity. AI systems can organise and summarise extensive case notes, assessments, reports, and policy documents, making salient details more visible and reducing reliance on fragmented records or memory. They can flag emerging risks or ethical concerns – such as patterns suggestive of cultural disadvantage or escalating safety issues – offering early prompts for reflection without prescribing specific outcomes. AI can also articulate multiple ethical perspectives on a dilemma, including duty-based, outcome-focused, and care-centred approaches, supporting richer supervision conversations and clearer recognition of trade-offs. In addition, AI systems can surface biased patterns in historical data, such as disproportionate "high-risk" classifications for particular groups, prompting practitioners to interrogate these patterns and exercise fairer judgement. In high-pressure environments such as intake, aged care, or mental health triage, AI may assist by identifying cases requiring rapid human review, reducing the likelihood that time-critical information is overlooked. Crucially, these benefits depend on practitioners retaining the authority to override system recommendations, with such overrides recognised as expressions of professional expertise rather than error. AI may also assist supervision by highlighting recurrent tension points—such as conflicts between rights and safety, or between client preferences and legal obligations—helping to slow decision cycles and encourage more deliberate, less reactive practice.

The ethical limitations of AI also reshape the demands placed on human services education. As algorithmic systems become embedded in assessment and decision-support processes, there is a risk that education shifts away from cultivating moral judgement toward training students primarily to operate and manage computational tools. Traditionally, human services education foregrounds historical, cultural, and political knowledge, including the history of Aboriginal and Torres Strait Islander peoples and the structural determinants of contemporary disadvantage. This knowledge is foundational, enabling practitioners to interpret situations ethically, recognise power asymmetries, and understand the historical forces shaping vulnerability (Ernst, 2024). In AI-mediated practice environments, however, such contextual understanding may be devalued if ethical reasoning is reframed as the correct configuration of inputs for algorithmic systems. The practitioner's role may drift from ethical agent to systems operator, tasked with supplying the "right variables" so that a system can generate a recommendation. This reorientation carries significant ethical risks: it implies that moral reasoning can be outsourced, that injustice is primarily a technical problem, and that relational knowledge is secondary to computational optimisation. Students may learn to encode vulnerability rather than recognise it, to defer to outputs rather than interrogate them, and to justify decisions procedurally rather than morally. If systems literacy comes to overshadow ethical literacy, human services education may inadvertently institutionalise the very "cruelly kind" harms this chapter seeks to caution against.

The hazards of over-reliance extend beyond pedagogy into the texture of everyday practice. Human services encompass child protection, elder care, mental health support, and addiction services – fields oriented toward justice, safety, and

compassionate social functioning. While some interventions must be rapid and highly structured, such as removing a child from immediate danger, these represent a narrow subset of professional activity. Much of human services work is relational, sustained, and contextually grounded, focusing on empowerment, trust-building, and support through complex life circumstances (Maguire Jack, 2015). When algorithmic systems are extended into this broader relational domain, practitioners may begin to treat automated recommendations as ethical reference points, narrowing the space for moral reflection, contextual sensitivity, and principled dissent. Empirical research underscores these risks. A 2024 meta-analysis of more than 60 studies reports declines of approximately 10–15% in public deductive reasoning, inference-making, and critical evaluation over three decades as automation has become more pervasive, alongside measurable slowdowns in creative output attributed to reliance on ready-made technological templates (George, Baskar, & Balaji, 2024). A 2025 study of 500 participants at Amity University in Haryana similarly found that prolonged AI use increased efficiency and consistency but diminished creative problem-solving, instinctual discernment, and critical thinking, reducing users' sense of agency as decision-making became increasingly shaped by automated systems (Shalu et al., 2025). Over time, AI systems may also entrench confirmation bias by narrowing the range of perspectives users encounter. In human services, this dynamic risks producing ethical fatigue, whereby practitioners disengage from moral reasoning because system-generated outcomes appear inevitable, eroding the sustained moral vigilance required in fields centred on vulnerability and power.

Finally, trust in human services depends on the belief that systems and institutions are capable of moral growth: acknowledging past failures, learning from harm, and adapting practice accordingly. Algorithmic systems, particularly those optimised for consistency, are poorly suited to such moral development. While AI can support judgement, it cannot experience harm, assume responsibility, or participate in ethical learning. As Joseph Weizenbaum argued, some acts of judgement ought to be attempted only by humans, because not all dimensions of care, responsibility, and moral reasoning are reducible to computational logic (Hasselberger, 2019). Ensuring that human judgement remains structurally indispensable is therefore not a nostalgic preference but a prerequisite for trust. Modernisation efforts that privilege efficiency over relational accountability may satisfy formal requirements while failing to meet the deeper needs, values, and concerns of service users (Smith, 2012). Over time, such approaches risk discouraging help-seeking, suppressing disclosure, and intensifying disengagement among communities already wary of institutional intervention. If deployed uncritically, algorithmic systems may therefore threaten not only individual wellbeing but the broader social legitimacy of human services.

Conclusion

Logic-based AI offers efficiency, consistency, and analytical support, yet it remains fundamentally limited within the emotionally, relationally, and morally complex

terrain of human services. It cannot possess moral agency or patiency, nor can it substitute for human judgement grounded in empathy, intuition, and contextual understanding. While specialised AI systems can assist ethical deliberation, the capacities required for decisions affecting life, dignity, and justice remain distinctively human. At the same time, AI is reshaping the moral environment by embedding assumptions into practice and redistributing responsibility. This dual reality – AI as indispensable support and as a potential source of "cruelly kind" outcomes – demands careful governance. Developers, institutions, and policymakers must prioritise fairness, transparency, contestability, and accountability across design and deployment, while ensuring that human practitioners retain normative authority over morally significant decisions. Only through robust, reflective, and human-centred oversight can AI enhance human services without compromising moral integrity or public trust (Mittelstadt et al., 2016; Floridi et al., 2018).

References

Atherton, D., Schwartz, R., Fontana, P., & Hall, P. (2023). *The language of trustworthy AI: An in-depth glossary of terms*. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-3

Baillergeau, E., & Duyvendak, J. W. (2016). Experiential knowledge as a resource for coping with uncertainty: Evidence and examples from the Netherlands. *Health, Risk & Society, 18*(7–8), 407–426. https://doi.org/10.1080/13698575.2016.1269878

Bryson, J. J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions* (pp. 63–74). John Benjamins.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15.

Dancy, J. (2001). Moral particularism. In Edward N. Zalta (Ed.). *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), https://plato.stanford.edu/archives/win2017/entries/moral-particularism

Donaher, K. (2021). Empathy in social care practice. *Journal of Social Work Values and Ethics, 18*(2), 45–57.

Ernst, J. S. (2024). Historical content in the social work curriculum: The value of local history. *Social Work Education, 43*(1), 174–185. https://doi.org/10.1080/02615479.2022.2098944

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

George, A. S., Baskar, T., & Balaji, P. (2024). Automation and cognitive decline: A meta-analysis. *Journal of Cognitive Studies, 42*(3), 233–251.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Hasselberger, W. (2019). Weizenbaum's critique of AI: Why some acts of thought should be human. *AI & Society, 34*(3), 501–509.

Hillman, K., & Chen, J. (2008). *Conflict resolution in end-of-life treatment decisions*. Sax Institute.

Inayatullah, S. (2001). The rights of a robot: Inclusion, courts and unexpected futures. *Journal of Futures Studies, 6*(2), 93–102.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kohlberg, L. (1981). *Essays on moral development: Vol. 1. The philosophy of moral development*. Harper & Row.

Maguire-Jack, K. (2015). Children's exposure to violence and the role of the social worker. *Child Abuse & Neglect, 48*, 1–10.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21.

NASW. (2021). *Code of ethics of the National Association of Social Workers*. NASW Press.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophy & Technology, 31*(4), 1–17.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

Resnik, D. B. (2017). The role of intuition in risk–benefit decision-making in human subjects research. *Accountability in Research, 24*(1), 1–29.

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Pearson.

Selbst, A. D., Boyd, D., Friedler, S., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta Georgia (pp. 59–68).

Shalu, R., Verma, N., Dev, K., Bhardwaj, A. B., & Kumar, M. (2025). Cognitive effects of prolonged AI use. *International Journal of Educational Technology, 19*(2), 112–128.

Shevlin, H. (2021). How could we know when a robot was a moral patient? *Cambridge Quarterly of Healthcare Ethics, 30*(3), 459–471.

Smith, C. (2012). Trust and the ethics of human service delivery. *Ethics and Social Welfare, 6*(3), 230–245.

Tronto, J. (1993). *Moral boundaries: A political argument for an ethic of care*. Routledge.

Veale, M., & Brass, I. (2019). Administration by algorithm? Public management meets public-sector machine learning. *Public Administration Review, 79*(5), 807–818.

14 Social Work Ethics and Generative AI

Michael Balkow

Introduction

Social work is a profession that works with people, in their homes, in hospitals with physical or mental health difficulties, seeing children in schools, and connecting with community and support networks in times when protection, support, advocacy, or empowerment is needed. Technology provides an interface between this human connection, from social worker to service user, to record, document, assess, and plan. These written social work records are vital but ultimately time-consuming, often starting off as handwritten notes to be expanded, edited, and inputted into systems and databases. Another key element of social work is the ability to think, to have the time and space to critically reflect upon work undertaken, whether through informal peer-to-peer discussion, or with more formal supervision routes, all of which require time, and critical thinking, often in fraught, emotionally charged contexts. It is perhaps unsurprising then that when technological advancement appears, which makes promises to increase productivity or reduce time-consuming administration tasks; that professions such as social work will take notice. Therefore, this chapter examines the use of generative AI in social work, the possibilities and potential, whilst paying attention to the risks, and the alteration of ethics of non-human entities when considering regulation. The scale and speed of technological advancements in AI can move at a rapid and unpredictably fast pace. The world of publishing and academic writing, with its need for diligent research, time for reflection and careful consideration, a peer review process, editing, and all the other seen and unseen facets of academia, makes AI a difficult topic to write about. What is written today may not be true tomorrow, so readers are advised to consider contexts. This is also a fundamental issue with AI, in that technology companies will be able to outrun anyone's abilities – whether organisations, the public, or governments – to carefully scrutinise or even regulate AI's usage and proliferation into our modern world. Social problems in society will endure, and social workers will remain to help solve these problems. However, what does change is the society around us, and social work must be malleable in a time of what Ioakimidis and Maglajlic (2025) refer to as 'fluid certainties.'

DOI: 10.4324/9781003627920-14
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

Generative Social Work

Generative AI is now gaining traction in social work practice in the United Kingdom, currently in the form of local authority trials, as 28 councils in England are using or testing an AI tool to produce case notes from visits and assessments. This has been developed by AI company Beam, alongside social workers, to record meetings and emails, with the AI then providing the user with a summary and suggested actions for inclusion in case notes based on council-agreed prompts (Koutsounia, 2024). Beam's website promises that you can 'Hire Self-Learning AI Agents to Run Your Operations,' boasting that their services are 'Used by Fortune 500 companies and scale-ups, the platform automates workflows with an advanced AI agent platform to reduce operational cost and create leverage' (Beam AI, 2025). Advertised on Microsoft's website is an article about their AI assistant Copilot and its usage in Barnsley, England where the anonymous 'Sam', a social worker enthuses:

> It's been really useful for writing up calls, tidying up notes or doing minutes after a meeting and making all that a lot quicker … It's really nice having Copilot as a sort of personal secretary … It means I'm not having to concentrate quite as much on scribbling things down, so it frees up brain space to really focus – if I'm chairing a meeting, for example.
>
> (Dawson, 2024)

The article then promotes further benefits and applications: 'The AI tool can quickly reorganise visit notes into accurate records … This saves social workers hours of analysis, so they can prioritise any urgent interventions. It can also quickly summarise large files when social workers take on a new case, presenting key information and major risk factors', followed by a rejoinder by Sam that: 'Given the importance of what we do, I do need to keep an oversight on it … But it makes a really useful start on things' (Dawson, 2024). The tone between Sam (the social worker), and the writer of the article, someone who works for Microsoft AI, bears some contradictions to the approaches. Sam is pleased to have a virtual secretary, to free up the space to really think and focus, whilst the author sees analysis of written case records as cumbersome, time-consuming and in need of AI to tidy up and organise these – presumably messy or disorganised – records, records that of course represent people's lives. Sam appears to value time to think about casework, whereas the author sees the reading of case records as unnecessary if AI can do this. We could argue that having AI transcribe a one-hour interprofessional core group meeting, as opposed to the social worker doing this, is of clear benefit to busy workers. However, critical questions arise from this: what could be lost as well as gained with this newfound efficiency? Turning notes from the one-hour meeting, into coherent minutes to share with professionals, may take time but could allow further insight, reflection and analysis, creating a more comprehensive understanding of the case. It could also follow that if extremely complex cases can be condensed and neatly summarised by AI, then there may be a temptation for managers to over-allocate workers or allocate these complex cases to less experienced workers.

Lillis, Leedham and Twiner (2020a, 2020b), drawing on their three-year ethnographically oriented study exploring contemporary professional social work writing, focused on a key concern – the amount of time taken up with writing, or 'paperwork.' This research framed social work as a 'writing-intense' profession, borrowing from Brandt who defines this as a profession where 30% of time is spent on writing (Brandt, 2005). The problem of paperwork distancing social workers from the service users and communities, has never been fully addressed or solved, although there have been some solutions proposed to make social work written recording part of its educational and professional standards (Balkow and Lillis, 2019). The AI tool also hints towards being more than a mere minute-taking secretary, into something that makes 'suggestions' possibly becoming a tool that will contribute to social work assessments. This is arguably a huge leap for social workers into an unknown future – the use of non-humans in decision-making which has potentially significant implications for the profession.

AI Futures

In 1997, Kreuger wrote a journal article titled: 'The End of Social Work', listing three factors that would lead to the demise of the profession in the next century. He defined one of these as the rise of 'Hypertechnologies', imagining a world where virtual headsets allow users to escape into enhanced realities, preferable to or improved upon their existing ones. Although today virtual reality headsets are not as ubiquitous as Kreuger anticipated, many young people spend several hours in the virtual space of social media, scrolling, clicking and liking hyper-real representations of self. Social media is an area of technology concurrent with the rise of the internet, where it can be debated the extent to which this harms the lives of young people, with psychologists like Jonathan Haidt proposing regulation and recommendations to minimise potential harm to children (Haidt, 2024). This is comparable in ways to the rise of generative AI. Just as it would be difficult to envisage a world where social media is totally removed from the lives of young people, we may soon arrive at a similar situation with AI in human-led workplaces.

It is easy to read the article today, with still at least seven decades until the end of social work, with an air of cynicism as some predictions appear far-fetched. However, the thought experiment is often a key part of literature in the development of AI and evaluating its societal impact, as exaggerated futures where AI becomes more intelligent that humans can be safely explored in the confines and safety of a book's pages (see Tegmark, 2017). The article displays some prescience with the prediction that counselling will be delivered by electronic software generated counsellors, seen in the rise of AI-powered mental health chatbots, replacing the face-to-face therapeutic relationship with an on-demand, readily available counsellor, of which multiple examples now exist (Woebot Health, 2025).

Kreuger also envisages how 'These developments will affect professional life … a vast array of new technologies will supersede human service skills as clients negotiate solutions in totally electronic personal territories' (Kreuger, 1997, p. 20).

The Covid-19 pandemic brought this to the fore in many workplaces including social work, as face-to-face visits and meetings, became wholesale conducted in the virtual spaces of Zoom, Teams or similar platforms. The residual effects of this being that many online meetings have remained, as with hybrid working models of work between home and the office (Ottmann and Noble, 2024; Ferguson et al., 2022). Kreuger's future social work is painted with gloomy pessimism, ignoring some of the strides technology can make in enhancing people's lives. Yet, by examining critical and negative perspectives of social work futures, we can interrogate ambitious or dubious claims and propose new possibilities. For example, another more recent article of a similar profession-apocalyptic tone is damning of social works' foundational knowledge base: 'With a divided focus, we are less grounded in evidence-based interpersonal interventions than psychology, less rigorous and sophisticated than sociology, less nimble and agile than genuine community movements, and less efficient than the algorithms that have come to dominate welfare bureaucracies' (Maylea, 2021, p. 778). Again, the tone sides with pessimism, discounting the value of social work having a fluid and eclectic theory base to draw upon, the possibilities of creativity and inclusion of varied knowledge, rather than a rigid and exclusionary approach. Yet this critique could present a possibility of AI being used to improve theoretical rigour in the profession, with generative AI quickly connecting research to practice. Evidence-based practice (EBP) can be defined as 'the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients' (Hannes, 2020, p. 1). This approach is dominant in the health professions and does have more difficulties when applied to mostly non-clinical professions like social work, as critiqued rigorously by Webb (2001). The evidence base from which EBP draws upon is to be found in research papers, the implication being that practitioners must be up to date with research at all times, utilising and then applying it to practice situations. Keeping professional knowledge current is vital but consumes time, drawing practitioners away from service users. As generative AI is trained on specific types of data that aligns with its purpose, an AI social work tool could be trained on huge volumes of social work literature, allowing practitioners to type in case scenarios, with generative AI then linking relevant literature, summarising studies and recommending outcomes based on findings. This process has already taken hold in the domain of health research, whereby every major consumer technology corporation, from Google to Apple, to Facebook, Amazon, Microsoft and IBM, has moved into the health and biomedical sector, referred to as the 'Googlization of health research' (Sharon, 2018). These companies hold the advantages of large-scale and broad reach, being skilled at handling large amounts of data, often for the purposes of selling this data to advertisers. Therefore, these companies have a utility here but may impart values of profit into health research (Filippo, 2024).

A similar but more refined approach is currently being used by the ethical philosopher Peter Singer, and his AI chatbot:

> Peter Singer AI is an advanced chatbot that facilitates meaningful ethical discussions. Drawing from my extensive body of work, it provides thoughtful

> responses and poses challenging questions to encourage users to reflect on their experiences and beliefs. Employing the Socratic method, this AI guides users through a process of critical thinking and self-examination.
>
> (Singer, 2024)

Users can propose ethical questions and the chatbot will emulate his responses. As Elliott (2022) identifies, chatbot's sound increasingly human-like, due to Natural Language Processing (NLP) which encompasses AI technologies related to the analysis, interpretation and the generation of language. However, NLP could be subject to bias either through the programming methods, or inherent bias contained within the data. For example, a study by Wilson, Hoffman and Morgenstern (2020), cited in (Kaplan, 2021, p. 27), illustrated that several decision-support systems applied by judges may be racially biased, based upon past rulings; and self-driving cars better detected lighter skin than darker tones, as their algorithms were trained using pictures among which were few people of colour. We have here a quite literal danger for people of certain races, if the technology is unable to correct for systemic racism within data. Therefore, before AI is even placed into the sphere of social work, questions would need to be asked about where the AI training data has been obtained from, does it reflect prevailing attitudes, norms, and therefore discrimination.

Generative AI Social Work Assessments

Anecdotally, from personal experience when having attended training sessions on how to integrate AI into the UK University curriculum, I am met with fatalist-type phrases in relation to AI such as: 'this stuff is going nowhere'. This is often a response to concerns around generative AI being used to plagiarise in student assignments, or concerns that all applications of AI and their subsequent outcomes, may not be entirely positive, such as effecting students' ability to write well academically and think critically. Generative AI gives us the ability to outsource our thinking, helping with idea generation, proofreading and critique of written work, functioning here like a critical friend or supervisor. Therefore, it would also be relevant to rephrase statements like these into: 'this stuff is going everywhere', AI being on every mobile phone, laptop, powering internet searches, in every workplace – every hospital, police station, court, and even placed into chips implanted in human brains. Case in point, Neuralink a firm founded by Elon Musk, which is developing plug in chips in human skulls to increase cognitive abilities and memory to 'save the human race' against AI (Dubber et al., 2021, p. 35). AI here being both the problem and the solution, an existential threat to humanity, yet one that can be surmounted with the fusion between soft brain matter, hard technology and ever more intelligent AI. While at this point AI remains fairly submissive and has not yet taken over the human population in some apocalyptic-style scenario, we should critically question which parts of society will benefit from AI to assist with human tasks, and to what extent AI should be everywhere or nowhere.

With this in consideration, I will explore the use of generative AI in social work assessments. Definitions of assessment have common themes such as ascertaining levels of risk and determining need (Milner et al., 2020). Bolton (2018) broadens the definition to include ethics and morals: 'Assessment is a complex moral and ethical act' (Bolton, 2018, p. 205). This considers assessment as more than just the sum of its parts – the mere gathering of data, written records, carefully organised and analysed with an appropriate outcome – but rather an act that incorporates philosophical concepts such as morals and ethics.

In a detailed book on social work assessments, Parker (2025) sees the enquiry following the tragic death of Victoria Climbie as an important point in the history of assessment in social work. Written by Lord Laming in 2003 (Laming, 2003), the report recommended action that social workers must undertake during the process of assessment including: 'speak alone to the child; see and speak with carers; visit the child's accommodation; seek and consider the views of other professionals' (Parker, 2025). These recommendations, which retrospectively can seem obvious and self-evident were not strictly adhered to at the time, the resulting outcome being the horrific torture and murder of an eight-year-old girl in the United Kingdom. We can note, that generative AI, in its current forms would not be able to undertake any of the tasks recommended by the review. This would require input from the social worker who would gather this primary data, generative AI could perceivably then assist with the written aspect of these assessments.

In addition to assessment definitions being broad, the steer has also altered historically. Munro (2008, 2020) discusses child and family assessments evolving to encompass risk: 'The language of risk is relatively new to child protection work. In childcare textbooks from the 1950s or 1960s, there was no reference to risk assessments' (Munro, 2008, p. 58). The term risk assessment can therefore be seen as rooted in prediction. If the risks to a child are identified then harm could feasibly be predicted, logically following that future harms could be prevented. Serious case reviews such as these lay bare the private family life and the individual suffering of children, exposing these to the arena of the public sphere. This often generates outcry, horror and a need for retribution against the perpetrators, frequently overshadowed and directed towards the professionals involved. A pertinent example in the UK being the case of Peter Connelly, or 'Baby P' as he was posthumously known (Jones, 2014). Government scrutiny and enquiry frequently follow with recommendations, and the need to fix practice errors could take a technological turn. Webb (2006) defines technology that enables social workers to do their job, as technologies of care. Predating AI, he lists these to include email and computerised case files, seeing them as tools but also entities that can be critically scrutinised:

> Technologies of care are concerned with the governance of uncertainty – they are rational schemas for organising social work and developing procedures for how it is carried out and delivered. In this sense they are programmatic since they are based on sets of calculating prescriptions through which care-based interventions are organised, arranged and regulated.
>
> (Webb, 2006, p. 141)

Webb further argues that care management 'tends to be overly preoccupied with the assessment of need and risk and the matching of available services for service users, rather than relationship building or therapeutic direct work' (Webb, 2006, p. 142). Technology here designed to assist, moves social work into a different direction, focusing on the collation and calculation of data to predict risk, then subsequently allocate services to meet need, rather than focusing on the individual relationship between social worker and service user. These technologies are still present, but may become increasingly enhanced or augmented by the addition of generative AI. For example, case files can be quickly summarised into shortened distillations of the most pertinent issues, such as trawling all associated notes regarding a parent's mental health history, and collating chronologies into a shortened summarised narrative. Chronologies are often a key part of social work assessments and can be critical in determining historical patterns, risk factors and providing vital links between disparate pieces of information. Dyke (2023) discusses that in practice, chronologies are often seen as a bureaucratic chore, rather than a vital part of understanding service users' history, recommending that they are the first thing a social worker should complete for an assessment. He also argues they have 'moral value' (Dyke, 2023, p. 33) reminding us that the service users' story matters.

To date, assessment in social work is quite rightly viewed through the lens of human-decision-making. Munro (2020, p. 27) discusses two defined methods of understanding the world as 'intuitive and analytic'. Intuition can be viewed as mysterious as it tends to happen at speed, an unconscious process that automatically occurs as a response to perceptions, which can then be articulated and justified by language. Drawing here from Nobel Prize-winning psychologist Daniel Kahneman (2011), he discusses an analogy of intuitive decision-making when firefighters entered a house with a kitchen on fire. Soon after entering the room, the commander shouted, without realising where this came from, to get out just before the floor collapsed. This was due to the fire being below rather than in the kitchen as originally thought. The firefighter discussed the decision being prompted by a 'sixth sense of danger' (Kahneman, 2011, p. 11). Psychologists now understand more about these types of intuitive decisions and there is less mystery involved. Seeing multiple similar situations over the course of his professional career will have contributed to a knowledge base the ability to seek similarities in patterns and situations. The need for firefighters to make quick intuitive decisions is vital, not having the luxury of time to weigh up, assess and analyse situations. However, we can apply similar principles to cases of child or adult protection in social work that require immediate responses. Generative AI's ability to read huge amounts of text, concisely summarising the key points could be applied to social work databases. These contain documentation collated by professionals and could be read by AI, then if also trained on relevant research could provide evidence-based responses and suggestions for outcomes. The speed at which this could happen would be impressive in comparison to human standards, however what is gained in speed and efficiency may be lost when information is glossed over in a cursory manner. For example, this could lead to what Ornellas et al. (2020), refer to as deprofessionalisation, a key

feature of neoliberalism which leads to reducing professional discretion, deskilling social work and diminishing professional autonomy and identity. Further warnings about generic tick-box approaches have been made by Dustin (2007), and the McDonaldization of social work, with efficiency, replication and productivity becoming the main measures of success, replacing core moral and ethical values of the profession.

Ethics and Generative AI Social Work

Banks (2021, p. 31) states that social work can be viewed as a 'human services' profession, an occupation that focuses on human relationships and the multiple complexities that arise therein. She identifies the relationship between social worker and service user as one of unequal power, the social worker holding more power than the service user, often the power to substantially intervene in that person's private family domain. If social work begins to use non-human entities in assessment and therefore decision-making, this could be viewed as a fundamental shift in the dynamics of the profession. This could distance social work from human elements in favour of AI-powered outcomes. To consider how generative AI may influence social work practice, and any subsequent alterations required to regulations and the ethics of the profession, we can examine the impact technology has had on social work historically. Writing in 2008, Parton, critically examined the use of computer databases within social work, something which is now ubiquitous in statutory social work. He argued that this created an environment where the key focus of activity, became the gathering, sharing and monitoring of information about individuals, together with the accounting for their own decisions and interventions, and those of the other associated professionals and agencies (Parton, 2008). The advantages of social care databases can be multiple, they allow information to be confidentially stored, easily and quickly accessed, then shared with other professionals. This allows continuity and adaptability of the workforce, in that if a social worker leaves post before the conclusion of their work, then another can take over the case, reading case notes and files to then continue to build upon the previous worker's role. However, Parton argues that:

> Rather than be concerned with presenting a picture of the subject, as, previously, social work increasingly acts to take subjects apart and then reassembles them according to the requirements of the database. Practitioners are required to produce dispersed and fragmented identities made up of a series of characteristics and pieces of information which are easy to input/output and compare. In the process, the embodied subject is in danger of disappearing and we are left with a variety of surface information … Such a development begs the question as to how much direct contact and interaction needs to take place between the social worker and the client if the primary concern becomes gathering, inputting and analysing information.
>
> (Parton, 2008, p. 263)

The focus of social work becomes less on the quality of social interaction with the service user, the human encounter, than on the amount of data that can be inputted into the database, signifying that work has been undertaken. In the example of the social worker leaving, this may be devastating for the service user who has taken time to establish trust with the worker, relaying painful details of their lives – fragmented and then be pieced back together. Yet, this is of no consequence to the database, which can continue to be inputted by the next worker – feeding the machine. Parton (2008) goes on to argue that this gives rise to a database culture within social work, moving away from narrative approaches where knowledge is constructed mainly from the face-to-face encounter between social worker and service user, to information and data collection, fitting the service users' biographical history into the narrow and defined parameters of the social care database, ultimately leading to the erosion of the social within social work. However, unlike the database, which is essentially a virtual space that holds data, generative AI can read at speed, rapidly creating summaries of data. In the same way that internet searches often now have an AI summary at the very beginning, generative AI could quickly read hundreds of case records relating to a service user, then summarise these into a narrative. The database with the active elements of AI then becomes something less inert – a mere storage warehouse of information, morphing into something that can think through data, mirroring the way a social worker would read case files, building a picture of the service users' life, to analyse patterns, identify risks, and needs. The time saving potential to social workers here is significant. Questions would however arise as to whether this time gained, would be converted to prioritising time spent with service users, versus working on administrative tasks, or whether this means quicker less diligent social work, as cases can be rushed through and completed at a much faster rate. Garrett, for example, warns us that: 'Clearly, there is a kernel of truth in the perception that social work is becoming more onerous and less fulfilling. The working day is expanded, and technology is, seemingly, deployed more to entrench the alienation of practitioners than to create possibilities for better service provision and personal growth' (Garrett, 2024, p. 124). Here technology is restricting rather than liberating the worker. The ability of generative AI to create space for social workers to reconnect with service users, away from the enduring administrative office tasks, may also create more focus and surveillance of populations as their data is gathered, stored and used by AI.

When we consider generative AI's role in social work practice, specifically if it takes a more active role in the assessment and decision-making process, then we may need to completely reconsider the ethical standards of the profession. Banks (2012) summarises that most of the literature on social work ethics has focused on lists of principles about how the social worker ought to treat the individual service user, often underpinned by one basic principle as 'respect for persons' (Banks, 2012, p. 42). The philosopher Immanuel Kant articulated this concept of the categorical imperative as: 'So act that you treat humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means' (Kant 1785, cited in Kerstein, 2024). Fundamental here is the idea of a

person, someone who has their own individual desires; ends which should not be subsumed by others for their own means to an end. This is comparative with the power that social workers can hold over others, and the need to protect persons from exploitation or abuse. The categorical imperative is not entirely perfect as an ethical standpoint, and there are situations whereby social workers would need to act for the common good. For example, a social work assessment that summarises a child should be removed from their parents, to protect them from future abuse, may go against the desires of that child and their parents, overriding their own ends for a specific means. However, present here are human persons and not robots, theoretical principles and guidelines of human-on-human behaviour. Therefore, when AI becomes involved in decision-making, social work ethics becomes blurred and problematic.

Benossi and Bernecker (2022) theoretically analyse robot ethics through a Kantian perspective. They argue that AI lacks the two key components required for morally relevant decision-making, which are an emotional inner life and freedom of will, clarifying that robots are not free, rather determined by preprogrammed algorithms. Then if AI lacks the core features both cognitively and ethically to make moral decisions, we can argue against their involvement in decision-making in social work, to risk diluting the ethical principles of this human profession. This would relegate generative AI usage to the function of an administrative assistant, leaving decision-making – which can be complex, emotionally challenging, mirroring the complexity of the problems to solve, amend or ameliorate – firmly in the hands of humans.

Conclusion

Responses towards AI can fall into different categories, from the sceptical to the optimistic, the fatalistic to the resistant. We are dealing with a phenomenon that is largely unencountered, the terrain is unknown and often the use of thought experiments, or guesswork is required to shape and predict AI futures. By analysing some of these responses we can move towards creating a critical framework to challenge assumptions, address risks and predictions of future workspaces. Failure to do this could result in retrospective action to reclaim areas of human practices, in areas that have become saturated with AI.

Elliott (2022, p. 10) states that, 'AI as both ubiquitous and invisible,' yet social work must not ignore or blindly accept its usage. The British Association of Social Workers (BASW) has issued a statement on the use of generative AI in social work, concluding that 'In light of current limitations, the report finds that the use of generative AI tools in care assessment and planning would currently be inappropriate without direct human oversight' (BASW, 2025, p. 11). This is a welcome warning but also vague in its stipulation. We do not know which human will have oversight and how direct this will be. Whether these will be managers, administrative workers, or people within the AI industry, the so-called Big Tech – the technology giants that dominate and control the market, often referred to by the acronyms GAFAM which refers to US-based companies and includes Google (Alphabet),

Apple, Facebook (Meta), Amazon, and Microsoft (Verdegem, 2024) – remains undetermined. The speed at which these companies develop if involved in social work, could overwhelm the professions' ability to regulate and define the ethical parameters of the work they undertake. However, if social work remains true to human-based ethics in all aspects of decision-making, then we can retain the most important elements – the human in this human profession.

References

Balkow, M., & Lillis, T. (2019). Social work writing and bureaucracy. *Citizen Network*. Retrieved May 29, 2025. https://citizen-network.org/library/social-work-writing-and-bureaucracy.html

Banks, S. (2012). *Ethics and values in social work* (4th ed.). Palgrave Macmillan.

Banks, S. (2021). *Ethics and values in social work* (5th ed.). Red Globe Press.

BASW (2025). The BASW statement on social work and generative artificial intelligence. Birmingham. Retrieved June 16, 2025. https://basw.co.uk/policy-and-practice/resources/basw-statement-social-work-and-generative-artificial-intelligence

Beam AI (2025). Self-Learning AI Agents for Enterprise Operations. Retrieved March 3, 2025. https://beam.ai/

Benossi, L. & Bernecker, S. (2022). A Kantian perspective on robot ethics. In Hyeongjoo Kim & Dieter Schönecker (Eds.). *Kant and Artificial Intelligence* [Online] (pp. 145–168). Germany: Walter de Gruyter GmbH.

Bolton, G. (2018). *Reflective practice: Writing and professional development* (5th ed.). SAGE Publications Ltd.

Brandt, D. (2005). Writing for a living: Literacy and the knowledge economy. *Written Communication, 22*(2), 166–197. https://doi.org/10.1177/0741088305275218

Dawson, A. (2024). How Barnsley Council is transforming local social care with AI. Retrieved March 3, 2025. https://news.microsoft.com/source/emea/features/how-barnsley-council-is-transforming-local-social-care-with-ai/

Dubber, M. D. et al. (eds.) (2021). *The Oxford handbook of ethics of AI*. Oxford University Press.

Dustin, D. (2007). *The McDonaldization of social work*. Ashgate. [Online].

Dyke, C. (2023). *Writing analytical assessments in social work* (3rd ed.). Routledge. https://doi.org/10.4324/9781041057970

Elliott, A. (2022). *The Routledge social science handbook of AI* (1st ed.). Routledge. https://doi.org/10.4324/9780429198533

Ferguson, H. et al. (2022). Social work and child protection for a post-pandemic world: The re-making of practice during COVID-19 and its renewal beyond it. *Journal of Social Work Practice*, *36*(1), 5–24.

Filippo, S. (2024). *Human freedom in the age of AI* (1st ed.). Taylor & Francis Group.

Garrett, P. M. (2024). *Social work and common sense: A critical examination*. Routledge.

Haidt, J. (2024). *The anxious generation: How the great rewiring of childhood is causing an epidemic of mental illness.* Allen Lane, an imprint of Penguin Books.

Hannes, K. (2020). *Evidence-based practice.* P. Atkinson et al. (Eds.). SAGE Publications Ltd.

Ioakimidis, V., & Maglajlic, R. (2025). Social work in the era of 'fluid certainties'. *The British Journal of Social Work*, *55*(1), 1–4. https://doi.org/10.1093/bjsw/bcaf021

Jones, R. (2014). *The story of baby P setting the record straight.* Policy Press.

Kahneman, D. (2011). *Thinking fast and slow.* Penguin Books.

Kaplan, A. (2021). Artificial Intelligence (AI): When humans and machines might have to coexist. In P. Verdegem (Ed.), *AI for everyone? Critical perspectives* (pp. 21–32). University of Westminster Press. https://doi.org/10.16997/book55.b

Kerstein, S. (2024). Treating persons as means. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Summer ed.). Retrieved April 30, 2025. https://plato.stanford.edu/archives/sum2024/entries/persons-means/

Koutsounia, A. (2024, October 4). AI could be time-saving for social workers but needs regulation, say sector bodies. *Community Care*. Retrieved March 3, 2025. https://www.communitycare.co.uk/2024/10/04/ai-could-be-time-saving-for-social-workers-but-needs-regulation-say-sector-bodies/

Kreuger, L. W. (1997). The end of social work. *Journal of Social Work Education*, *33*(1), 19–27.

Laming, L. (2003). The Victoria Climbie Inquiry: Report of an inquiry by Lord Laming. Retrieved March 18, 2025. https://www.gov.uk/government/publications/the-victoria-climbie-inquiry-report-of-an-inquiry-by-lord-laming

Lillis, T., Leedham, M., & Twiner, A. (2020a). Time, the written record, and professional practice: The case of contemporary social work. *Written Communication*, *37*(4), 431–486. https://doi.org/10.1177/0741088320938804

Lillis, T., Leedham, M., & Twiner, A. (2020b). 'If it's not written down it didn't happen': Contemporary social work as a writing intensive profession. *Journal of Applied Linguistics and Professional Practice*, *14*(1), 29–52. https://utppublishing.com/doi/10.1558/jalpp.36377

Maylea, C. (2021). The end of social work. *The British Journal of Social Work.* [Online] *51*(2), 772–789. https://academic.oup.com/bjsw/article-abstract/51/2/772/6018482

Milner, J., Myers, S., & O'Byrne, P. (2020). *Assessment in Social Work.* (5th ed.). Red Globe Press.

Munro, E. (2008). *Effective child protection* (2nd ed.). Sage Publications Ltd.

Munro, E. (2020). *Effective child protection* (3rd ed.). Sage Publications Ltd. https://doi.org/10.1080/02650533.2018.1448768

Ornellas, A. et al. (2020). The Fourfold Neoliberal Impact on social work and why this matters in times of the COVID-19 Pandemic and beyond. *Social Work, 56*(3), 235–249. https://doi.org/10.15270/56-4-854

Ottmann, G., & Noble, C. (2024). *Post-pandemic welfare and social work: Re-imagining the new normal* (1st ed.). Routledge. https://doi.org/10.4324/9781003416210

Parker, J. (2025). *Social work practice: Assessment, planning, intervention & review* (7th ed.). Jonathan Parker. Learning Matters.

Parton, N. (2008). Changes in the form of knowledge in social work: From the 'social' to the 'informational'? *The British Journal of Social Work, 38*(2), 253–269. https://doi.org/10.1093/bjsw/bcl337

Sharon, T. (2018). When digital health meets digital capitalism, how many common goods are at stake? *Big Data & Society*, *5*(2). https://doi.org/10.1177/2053951718819032

Singer, P. (2024). Introducing peter singer AI: Elevating ethical discourse in the digital age. Retrieved May 2, 2025. https://boldreasoningwithpetersinger.substack.com/p/introducing-peter-singer-ai-elevating-f17

Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence.* Penguin Books.

Verdegem, P. (2024). Dismantling AI capitalism: The commons as an alternative to the power concentration of Big Tech. *AI & Society*, *39*(2), 727–737. https://link.springer.com/article/10.1007/s00146-022-01437-8

Webb, S. A. (2001). Some considerations on the validity of evidence-based practice in social work. *The British Journal of Social Work, 31*(1), 57–79. https://doi.org/10.1093/bjsw/31.1.57

Webb, S. A. (2006). *Social work in a risk society: Social and political perspectives*. Palgrave Macmillan.

Wilson, B., Hoffman, J., & Morgenstern, J. (2020). Predictive Inquiry in Object Detection. Working paper.

Woebot Health (2025). Mental health needs have multiplied. Support hasn't. Until now. Retrieved February 28, 2025. https://woebothealth.com/

15 Virtual Encounters in Bereavement Care

Ethical and Clinical Perspectives on AI and Virtual Reality

Seounghoon Oh, Ranaa Almasoudi, and Catriona Mayland

Introduction

This chapter critically examines the growing incorporation of virtual reality (VR) and artificial intelligence (AI) into bereavement care, not as a technological solution to grief, but as an ethical and relational challenge. As immersive technologies increasingly enter palliative and bereavement settings, they raise urgent questions about how grief, memory, and relational responsibility are mediated through technological encounters. Against this backdrop, the chapter advances a central argument: that VR and AI in bereavement care should not be oriented toward perpetual reunion or ongoing presence, but designed as bounded, time-limited rituals that support containment, meaning-making, and return to everyday life.

While such technologies may offer families opportunities to express what was left unsaid and to sustain continuing bonds (Klass, 1996; Stroebe & Schut, 1999; Slater & Sanchez-Vives, 2016), they may also blur the boundary between memory and simulation, particularly when family members hold different memories or participate in these encounters in unequal ways (Madary & Metzinger, 2016). We argue that VR and AI in bereavement care should function as a bridge back to life facilitating meaning-making and the accommodation of loss (Neimeyer & Thompson, 2014)—rather than a destination.

Using virtual ethnography to analyse the VR-based documentary series produced by Munhwa Broadcasting Corporation's (MBC)'s 'I Met You' (Seasons 1–4, 2020–2024) as a situated cultural practice—characterised by emotionally intense reunions, uneven family participation, scripted farewells, and public mediation of private grief (Hine, 2000; Boellstorff et al., 2012)—this chapter develops a practice-oriented ethical framework for palliative care practitioners working within multidisciplinary teams. The framework comprises five interlinked dimensions: (I) emotional and psychological safety; (II) relational ethics; (III) ritual design; (IV) boundaries of representation—ethics and safeguards; and (V) memory and social justice. Each dimension maps onto concrete professional tasks—briefing and consent, family mediation, trauma-informed scripting, careful data governance, structured closure and follow-up—that we detail in the body of the chapter.

Methodologically, we exclude VR imagery to avoid directing attention toward the immersive visual features of the technology and instead invite readers to engage

DOI: 10.4324/9781003627920-15

This chapter has been made available under a CC-BY-NC-ND 4.0 license.

with the narrative and relational aspects of grief. This aligns with Polanyi's framework; just as palliative care practitioners should discern tacit dimensions hidden beneath the surface, we invite the reader to look beyond the explicit image. By removing the visual, we are challenged to 'indwell' in the narrative and perceive the human dynamics that technology cannot capture.

Theoretical Framework: Tacit Knowledge and the Gestalt of Grief

This chapter draws on Michael Polanyi's concept of tacit knowledge to frame its five ethical dimensions (Polanyi, 1958, 1966). In VR and AI-enhanced bereavement care, ethical judgement cannot rely on protocols or consent forms alone. It also depends on practitioners' attention to subtle, embodied cues—such as hesitation, silence, or discomfort—that shape how encounters are actually experienced.

From this perspective, safety and ritual design are best understood as a *Gestalt*: not a checklist of discrete steps, but a whole that emerges from how procedures, relationships, and emotional cues are held together in practice. While AI systems can follow explicit scripts, they cannot recognise or respond to the tacit weight of grief. Human judgement therefore remains essential to ensure that VR encounters remain contained, meaningful, and ethically responsive.

This distinction becomes particularly salient with the advent of generative AI, where autonomy and unpredictability constitute a central ethical disruption. Unlike scripted VR encounters, generative AI systems can produce responses that are not fully anticipated or constrained. In this regard, Polanyi's (1966) insight is instructive: while machines are capable of processing explicit rules, they lack the tacit coefficient of personal commitment that underpins responsible judgement. An AI avatar may simulate human speech, but it cannot dwell in the shared reality of loss or bear responsibility for the moral consequences of its utterances (Polanyi & Prosch, 1975).

From this theoretical perspective, the notion of the palliative care practitioner as a 'human circuit breaker' can be understood as an analytic concept that captures how tacit professional judgement may become salient when AI-mediated encounters exceed or destabilise ethically negotiated boundaries. Rather than prescribing a fixed intervention, this concept highlights the limits of algorithmic autonomy and the continuing need for human responsibility in contexts of grief, vulnerability, and moral risk.

Applying this theoretical lens allows us to move beyond technological spectacle and attend to the subtle human dynamics through which ethical boundaries are established, negotiated, and sometimes breached. The following section introduces the empirical case study that grounds this analysis and illustrates how these dynamics unfold in practice.

MBC's Meeting You (Seasons 1–4): Nine Encounters in Grief and Farewell

As a public-service documentary series, I Met You (너를 만나다), produced by MBC, offers a distinctive case for examining VR-mediated bereavement encounters

through virtual ethnography (Boellstorff et al., 2012). Across its 2020–2024 seasons, the programme typically follows one bereaved family per episode through a structured process: initial interviews with family members; the co-design of a virtual encounter using personal photographs, home videos, and recorded memories; and a final, time-limited VR session staged in a controlled studio environment. During this session, one participant—such as a bereaved parent, bereaved spouse, or an adult child mourning a deceased parent—wears a VR headset and is guided through an immersive scene in which a digitally rendered representation of the deceased appears. The encounter is usually scaffolded by a pre-agreed script and facilitated prompts, culminating in an explicit farewell. Other family members may be present either in the same space or nearby, and the production team's mediation—through framing, editing, and emotional pacing—forms part of the encounter's conditions.

For the purposes of this chapter, the series is not treated as a clinical intervention to be replicated. Rather, it is analysed as a culturally mediated portrayal of grief that makes visible recurring ethical tensions: who participates and who is excluded; how families negotiate competing memories and meanings; how scripted 'closure' may comfort some while unsettling others; and how public representation of private grief can intensify vulnerability. This section analyses nine encounters to ground the chapter's ethical framework and to draw out practice-relevant lessons for palliative social work within multidisciplinary palliative care settings (Table 15.1).

Five Dimensions of Palliative Care Social Work in VR/AI-Enhanced Bereavement Care

As virtual ethnographers, we examined MBC's *I Met You* (Seasons 1–4), which staged nine VR/AI encounters of grief and farewell. The cases offer lessons for practice, showing how palliative care settings can use these experiences to guide ethically responsible applications of VR/AI in bereavement care. They highlight five interrelated dimensions: emotional and psychological safety, relational ethics, ritual design, boundaries of representation, and memory in relation to structural injustice.

Dimension I: Emotional and Psychological Safety

Only in Season 4 were psychologists involved, and even then, the broader remit of psychosocial coordination—typically fulfilled by one of the palliative care practitioners—remained absent. This is significant because emotional and psychological safety is foundational to bereavement interventions (Reith & Payne, 2009). While Korean law designates social workers as essential members of hospice and palliative care teams, such statutory arrangements are not universal. Internationally, roles in VR/AI-enhanced bereavement care are shaped by local service configurations and are enacted by palliative care practitioners across disciplines.

Analysis of VR- and AI-mediated bereavement encounters indicates that psychosocial and ethical vulnerabilities arise across three phases: preparation, the encounter itself, and aftercare. Preparation involves psychosocial assessment,

Table 15.1 Case summaries of VR bereavement encounters in meeting you

Episode (thematically grouped)	*VR participants*	*VR encounter design*	*Key themes*	*Observed gap in bereavement care practice*	*Corresponding ethical dimension(s)*	*Practice signal*
1–2	Mother–Daughter (VR)	Birthday ritual with chosen dress, cake, song; siblings present or observing	VR birthday encounter with reassurance ("I'm not in pain"), gratitude, goodbye	High emotional intensity combined with immersive realism, without clearly articulated containment or post-encounter support	Dimension I – Emotional and Psychological Safety Dimension II – Relational Ethics Dimension III – Ritual Design	Time-limited rituals for unfinished farewells; offer siblings participation options to avoid coercion
3–4	Husband–Wife (VR)	Weeks of preparation; favourite locations; scripted humour and dance	VR reunion as 'date' with dancing, closure	Extended preparation and scripted intimacy risked reinforcing idealised relational narratives, with limited attention to boundaries that prevent prolonged emotional dependency	Dimension I – Emotional and Psychological Safety Dimension III – Ritual Design Dimension IV – Boundaries of Representation	Reframe as ritual (opening–peak–closure); minimise cloning, signpost touch to prevent serial dependency

(*Continued*)

Table 15.1 (Continued)

Episode (thematically grouped)	*VR participants*	*VR encounter design*	*Key themes*	*Observed gap in bereavement care practice*	*Corresponding ethical dimension(s)*	*Practice signal*
6–7	Mother (VR) and Family	Everyday practices (recipe, cup, gardening); Symbolic gestures; garden; coached language; explicit exit and follow-up	Legacy, continuing bonds in daily life Closure, reassurance, emotional healing	Everyday symbolic practices were meaningfully integrated, yet responsibilities for coordinating family participation and ensuring structured closure were not explicitly supported	Dimension II – Relational Ethics Dimension III – Ritual Design	Ordinary remembrance can be more sustainable bonds via routines of VR, craft exits and aftercare for management
9	Son (VR), Mother and Father	Limited interaction with slightly aged-up avatar; symbolic space (bench, tree)	Legacy, growth, risk of simulated continuity	Ambiguity around developmental boundaries and representational limits heightened the risk of simulated continuity, particularly in the absence of clear constraints on interactivity and duration	Dimension I – Emotional and Psychological Safety Dimension IV – Boundaries of Representation	Bound interactivity and time; avoid persistent or 'living' avatars

(*Continued*)

Table 15.1 (Continued)

Episode (thematically grouped)	*VR participants*	*VR encounter design*	*Key themes*	*Observed gap in bereavement care practice*	*Corresponding ethical dimension(s)*	*Practice signal*
5 Social injustice	Kim (VR) Occupational death and Citizen	Non-reunion; narrative tracing preventable death; public call to action	Witness, injustice, policy petition	Grief was mobilised toward public testimony and advocacy, yet limited safeguards were evident to protect families from overexposure, retraumatisation, or instrumentalisation of loss	Dimension IV – Boundaries of Representation Dimension V – Memory and Social Justice	Link grief to advocacy; shield families from overexposure or retraumatisation
8Social sacrifice	Firefighter (public civic death VR) and Citizen	VR tribute with opening dedication, civic tone, graded intensity, firm closing	Public memory, honour, safety culture	Commemorative framing prioritised collective honour, while mechanisms to differentiate mourning from public spectacle and to support bereaved families privately were underdeveloped	Dimension III – Ritual Design Dimension V – Memory and Social Justice	Avoid gamification; separate commemoration from operational functions

risk screening, and clear communication of participants' rights and the limits of the encounter. During the encounter, safeguarding depends on human facilitation to uphold boundaries and to implement pre-agreed pause or exit procedures in response to emotional escalation (Bailenson, 2018). After the encounter, the absence of structured debriefing and follow-up may leave participants without adequate support to integrate the experience (SAMHSA, 2014).

In the context of generative AI, safety cannot be understood as a technical checklist alone. AI-driven avatars introduce unpredictability, increasing the risk of emotional overload if encounters are insufficiently contained. This underscores the importance of human oversight capable of recognising and responding to emerging distress. Within multidisciplinary palliative care practice, social work contributes particular expertise in psychosocial assessment, relational mediation, and ethical boundary-setting, alongside clinical and nursing roles (Reith & Payne, 2009).

The documentaries vividly illustrate why safety should be prioritised. In Season 1, Jang Ji-sung, who lost her seven-year-old daughter Nayeon, was reunited with her in VR (MBC, 2020). The reunion culminated in a simulated birthday party—a scene combining joy, longing, and grief. Although VR encounters powerfully evoked continuing bonds (Klass, 1996), analysis across all nine sessions indicated that emotional intensity became ethically significant only when combined with immersive VR, limited participant control, and insufficient aftercare (Worden, 2018). The aftermath also highlighted that potential harms may extend beyond the VR session itself. In this publicly mediated case, Jang was exposed to intense online scrutiny and hostility, illustrating how secondary distress may arise when immersive bereavement encounters are broadcast rather than privately contained. While this risk is specific to public settings, it points to a broader gap in how bereavement care has addressed VR- and AI-mediated encounters. In the absence of established multidisciplinary frameworks, responsibilities for anticipating, containing, and responding to such harms remain unclear, underscoring the need for closer clinical and cross-disciplinary consideration of how bereavement care principles should be adapted in these emerging contexts.

Dimension II: Relational Ethics

Relational ethics in VR bereavement care concerns who participates, on what terms, and with what safeguards. Families bring grief histories, traditions, and power dynamics that can support or harm. Rather than reflecting the failure of any single profession, this absence highlights a broader gap in how bereavement care has approached immersive technologies, underscoring the need for clearer multidisciplinary integration before roles can be specified. *I Met You* foregrounded reconciliation but obscured conflict, underscoring how family relations—particularly in collectivist contexts—may be shaped by relational paternalism, where actions framed as care or protection can inadvertently suppress disagreement and complexity. These dynamics raise ethical tensions around autonomy, consent, and boundary-setting that cannot be resolved through formal procedures alone. Agreement may be expressed, yet shaped by unspoken pressure or relational obligation.

Ethical attention is therefore required not only to what is said, but to how participation is negotiated, experienced, and sustained in practice.

Beyond One-Off Consent: Autonomy in Family-Oriented Systems

Consent in VR and AI-enhanced bereavement care should not be treated as a one-off formality but as an ongoing relational process (Banks, 2020). The seductive realism of generative AI creates a 'technological imperative,' subtly constraining the perceived ability to refuse or withdraw. In this context, palliative care practitioners should verify understanding of risks, ensure that children and young people have genuine options to assent or decline, and protect participants from subtle forms of coercion (Alderson & Morrow, 2020). In practice, this requires plain-language explanations and embedding 'micro-consent' prompts like 'Shall we continue?' throughout the session.

I Met You illustrates the ethical risks when safeguards are unclear. In Season 3, a daughter reunited with her deceased mother alongside her family, yet the programme gave little sense of her own right to decline. In publicly mediated settings, collective expectations may shape participation in ways that complicate individual autonomy. From a palliative care perspective, this underscores the need for shared ethical safeguards—such as layered consent and clear withdrawal mechanisms—rather than reliance on any single professional role.

In family-oriented systems such as Korea, ethical risks often arise from forms of *relational paternalism*, where decisions are guided by perceived collective care rather than individual preference. In these contexts, family members may assume that participation in rituals is inherently beneficial, especially when framed as acts of love, remembrance, or healing. Such paternalistic dynamics can unintentionally silence dissent, particularly where refusal is interpreted as disloyalty or emotional withdrawal. As a result, vulnerable members—especially children—may feel obliged to participate in high-intensity VR/AI encounters in order to preserve family harmony, even when this participation compromises their own psychological safety.

As discussed earlier through Polanyi's concept of tacit knowledge, the palliative care practitioner's task is to detect the tacit signs of this coercion. In a high-context culture, a refusal is rarely voiced explicitly. Pressure may be masked by the collective atmosphere, and resistance surfaces only in 'tacit' cues—silences, reluctant gestures, averted eyes, or deferred answers. Here, the practitioner functions as a protective filter: reading these embodied signals that contradict the signed consent form. In such situations, the practitioner should interrupt the process gently but firmly—reminding participants that 'honouring the dead does not require sacrificing the living'—thereby ensuring that autonomy remains genuine rather than performative.

Family Mediation and Participation

VR- and AI-mediated bereavement encounters are inherently social events, shaped by family participation, cultural expectations, and shared rituals rather than solitary engagement alone. In the lead author's ethnographic reading of *I Met You*,

grief unfolded within family systems where roles, expectations, and unresolved conflicts shaped how loss was lived. Palliative care literature consistently identifies the family as the primary unit of care, requiring active mediation to safeguard both individuals and relationships (Firth, Luff & Oliviere, 2005).

The absence of such mediation was striking in Season 3, where a daughter's reunion with her deceased mother provoked unmanaged collective sobbing. Here, the AI-driven spectacle amplified emotion but failed to contain it. From this perspective, what becomes visible is the need for careful structuring of participation, attentive to tacit emotional cues such as hesitation, silence, or relational strain. Without such attention, immersive encounters risk intensifying unresolved family dynamics rather than containing grief. Crucially, this involves establishing a 'ritual charter' to set ground rules for respectful speech, short turns, and explicit exit rights, ensuring the VR space does not reproduce hidden hierarchies (Hudson & O'Connor, 2007).

Most importantly, participation must be diversified to counter the binary logic of technology (user vs. non-user). Allocating graded roles allows family members to engage at their own pace: some may wear the headset (active participants), while others act as 'witnesses,' 'guardians' of memory objects, or 'supporters' holding hands. This spectrum of engagement affirms agency, particularly for children who may feel pressured to conform (Alderson & Morrow, 2020).

Dimension III: Ritual Design

VR/AI bereavement encounters do not replicate reality; they reconfigure memories into ritualised acts of farewell. However, the therapeutic power lies not merely in the final VR output, but in the preparatory process of co-creation. Families bring cherished experiences—birthdays, meals, songs—that VR/AI amplifies into symbolic forms. This selection process is profoundly significant: it requires the family to sit together, sift through decades of memory, and agree on 'what matters most.' This resonates with Turner's (1969) account of ritual as the transformation of ordinary acts into symbolic meaning. Without this intentional, collaborative design, encounters risk becoming trivial spectacles.

Applying the theoretical framework, ritual design operates as a *Gestalt*, where the practitioner helps the family integrate explicit elements (scripts, objects) with implicit meanings. The palliative care practitioner's role is to facilitate this 'narrative distillation': guiding the family to choose a single, shared symbol—a specific dish, a favourite song—that can anchor their collective grief. This participatory process ensures that the VR/AI encounter is not a solitary hallucination for one user, but a shared cultural text that every family member has helped to write.

The documentaries illustrate the risks when this focused design is absent. As noted regarding Season 1, layering multiple symbolic peaks-such as a simulated birthday cake, a favourite song, and repeated virtual embraces -created emotional flooding without containment. Ritual theory suggests ceremonies are most effective when centred on a single, bounded act of reassurance (Walter, 2015). Conversely, Season 2 highlighted the power of distilled symbolism. A husband encountering

his wife cooking a familiar dish found comfort in the symbolic gesture rather than photorealistic perfection. The design process, therefore, moves from 'replicating everything' to 'selecting the essential,' a task that requires the palliative care practitioner's phronetic guidance. This provides contextual grounding for understanding how expertise in mediation and safeguarding may inform the co-design of ritualised VR encounters, without positioning any single profession as exclusively responsible.

Seen in this way, ritual design is not only about symbolic selection but about structuring participation. Without such design, VR encounters tend to recentre grief around a single 'experiencer', inadvertently marginalising family members who remain outside the headset. A focused design process anticipates this risk by distributing meaningful roles across the family system—for example, through shared decisions (Berghs et al., 2022) about scripts, music, or the framing of the farewell. In this form, ritual design helps transform VR from an isolated, individual experience into a bounded family ritual. By clarifying who participates, how, and with what limits, it supports emotional containment while preserving the collective nature of mourning.

Dimension IV: Boundaries of Representation: Ethics and Safeguards

Representation in VR- and AI-enhanced bereavement encounters is fraught with ethical dilemmas, raising questions about authenticity, ownership, consent, and the potential for harm. To reconstruct a deceased family member's face, voice, or gestures raises fundamental questions about authenticity, ownership, and potential harm. While *I Met You* illustrated the emotional resonance of such encounters, it revealed little about how avatars were sourced or retired. In therapeutic practice, however, representation acts are inseparable from informed consent, data governance, and symbolic limits (Buitelaar, 2017).

A critical and often underexamined ethical issue in VR- and AI-enhanced bereavement care concerns the negotiation of the public–private boundary. Immersive technologies can blur this boundary by transforming intimate grief into shareable or reproducible content, raising questions about exposure, consent, and secondary harm beyond the encounter itself. While such risks may be amplified in publicly mediated contexts, they are not limited to broadcast settings. The *I Met You* case illustrates this dynamic in an intensified form: following the televised reunion with her daughter, Nayeon's mother was subjected to sustained public scrutiny and hostility, eventually closing her personal blog in response (Lee, 2020). This case vividly illustrates how comfort experienced *within* the VR/AI session can be undone by exposure *outside* it. Representation, therefore, requires active negotiation of the 'ethics of witnessing'—determining who is permitted to see the dead and under what conditions (Walter, 2015).

More broadly, representation is a site of power and vulnerability (Hall, 1997). Without clear safeguards, avatars risk becoming what we term 'unbounded representations': digital forms that persist beyond their intended ritual context (Ramirez & LaBarge, 2018). This risk intensifies with generative AI, which introduces the

danger of 'hallucination'—fabricating words the deceased never spoke. Such fabrication constitutes a form of 'narrative violence' that contaminates the authentic memories of the bereaved (Hollanek & Nowaczyk-Basińska, 2024). Ethical practice should therefore respect the deceased as a figure of 'memory' (fixed in the past) rather than allowing them to perform as a malleable 'simulation' (evolving in the present). Without this distinction, families face tangible risks: impersonation, fraud, stalking, child protection violations, defamation, data leakage, and commercial exploitation (Stokes, 2012; Brubaker, Hayes & Dourish, 2013).

These are not abstract possibilities, but anxieties voiced by families themselves. In *I Met You*, parents described the pain of 'suspended time,' fearing that their child was 'frozen' while the world moved on. Such testimonies reveal tacit fears that digital continuity may haunt rather than console (Brubaker, Hayes & Dourish, 2013; Buitelaar, 2017). Here, Polanyi's framework is instructive: representation is not only about explicit code but also about tacit contexts of reception—shame, pride, and vulnerability—that silently shape how families experience being seen (Polanyi, 1966).

To mitigate these risks, palliative care practitioners should enforce boundaries that are both explicit and tacit. Explicitly, avatars should remain symbolic and finite: confined to single-use encounters and subject to strict deletion protocols (Floridi, 2013). Any suggestion of autonomous growth or perpetual dialogue—features often marketed by AI companies—should be excluded to prevent ethical confusion (Öhman & Floridi, 2017; Hollanek & Nowaczyk-Basińska, 2024). Tacitly, practitioners should attend to embodied cues like hesitation or posture that indicate emotional saturation. As noted in the theoretical framework, safety emerges as a *Gestalt*, where explicit safeguards merge with this tacit discernment (Polanyi, 1966).

Fundamentally, safeguards open space for symbolism. Turner (1969) reminds us that rituals derive power from symbolic gestures rather than literal replication. Just as legacy letters provide comfort, VR silhouettes or partial voice fragments may evoke presence more effectively than photorealistic likenesses, precisely because they leave room for interpretation (Chochinov et al., 2005). In this way, palliative care practitioners mediate what should remain absent, ensuring that digital encounters remain bounded, symbolic, and oriented toward healing (Walter, 2015).

Dimension V: Memory and Social Justice

Memory and legacy form the deepest layer of bereavement care, shaping not only how the dead are remembered but also how the living carry remembrance forward. Families in *I Met You* frequently articulated a sense of time suspended. In Episode 2, a mother reflected, "It feels as though my child's time has stopped," revealing memory as both comfort and burden. These practices echo bereavement traditions across cultures, where legacy work—letters, recipes, stories—serves as continuity, weaving remembrance into daily life (Chochinov et al., 2005). Immersive technologies intensify these possibilities, yet their value lies not in replication but in helping families translate memory into symbolic forms.

The convergence of VR and AI expands legacy work, offering tools such as curated archives or 'legacy capsules.' However, these innovations raise ethical anxieties. In the AI era, there is a profound risk that families remain anchored to the past, blocking their resilience and the enjoyment of a new life. Instead of fostering a return to the present, unrestricted interaction with a digital avatar can paradoxically weaken the solidarity of the living, disrupting the formation of strong bonds among remaining family members who should navigate their future together. This risk echoes the danger of 'unbounded representation' discussed in Dimension IV, where likenesses persist beyond ritual use, creating dependency (Ramirez & LaBarge, 2018). Fears that an avatar might speak autonomously reflect wider concerns about digital remains haunting the living (Brubaker, Hayes & Dourish, 2013; Buitelaar, 2017). To safeguard families, AI design should avoid autonomous growth. Instead, it should focus on symbolic artefacts—finite, modest, respectful—that preserve memory while acknowledging absence, aligning with information ethics (Floridi, 2013) and memorialisation studies (Walter, 2015).

However, if AI risks trapping memory in the past, social justice transforms memory into a force for the future. In Episode 4, families bereaved by an industrial disaster used VR to retrace their children's final steps, transforming grief into indignation: "My son should never have died at work." In Episode 8, firefighters' families pleaded, "Do not let their sacrifice be in vain." Here mourning became testimony, exposing unsafe labour practices. Such encounters show how remembrance can transcend consolation to become civic testimony, linking private grief with public critique (Holloway & Taplin, 2013).

From a Polanyian perspective, this duality reflects the tension between *commitment* and *calling*. Polanyi argued that knowledge requires *commitment*—a personal stake in holding something true—and *calling*—an orientation toward responsibility beyond oneself (Polanyi, 1958, 1966). Families' intimate acts of remembrance exemplify commitment, while testimonies emerging in contexts of industrial or social disaster reflect calling: grief extending outward to demand recognition of structural injustice, accountability, and the acknowledgement of harms that have been historically silenced. Bereavement care thus becomes a dual practice—safeguarding symbolic legacies within families while also supporting forms of remembrance that give public voice to otherwise unseen victims and link grief to broader claims for social justice.

Palliative care practitioners play a pivotal role in enabling this practice. They help families identify symbolic anchors that can be safely integrated into rituals and connect them to networks of solidarity (BASW, 2016). Their role requires explicit safeguards—ensuring avatars are finite—and tacit discernment, reading silences that signal when memory work risks harm. Thus, Dimension V highlights a vital insight: to remember well is not only to preserve personal meaning, but also to contribute to forms of justice understood here as recognition, accountability, and the public acknowledgement of harms that have been historically overlooked or silenced. Across the cases and theoretical analyses discussed in Section 4.5, VR- and AI-mediated remembrance is shown to be most ethically viable when it remains finite, symbolic, and safely bounded—conditions that help prevent

emotional flooding, preserve agency, and support integration after the encounter. At the same time, such bounded remembrance can still amplify testimony by giving structured, public form to experiences of loss, particularly in contexts of social or industrial harm. Families' words—"This should not happen again"—speak to the preventable social and industrial harms underlying the loss, illustrating how private memory can extend into calls for public accountability and reform. For palliative care practitioners, the challenge is to support this movement: honouring the dignity of memory while supporting callings for justice understood as recognition, accountability, and redress for preventable social and industrial harms that gave rise to the loss (Banks, 2020; Walter, 2015).

Limitations and Implications

The *I Met You* series reveals both the promise and the limits of immersive technologies in bereavement care. As a broadcast production, it highlighted reconciliation but left little space for conflict or fractured realities, risking a unified narrative that overlooks the complexity of grief (Walter, 2015). My analysis is further constrained by the observational nature of virtual ethnography and the cultural specificity of the Korean context, which may not translate directly elsewhere (Oh, 2025). Added to this is the challenge of technological volatility: frameworks proposed today require constant vigilance as generative AI evolves from scripted VR to autonomous avatars, introducing new risks of unpredictability and hallucination (Floridi, 2013).

Importantly, the chapter has not examined technologies oriented toward ongoing conversational engagement with the dead, such as so-called 'death chatbots'. These systems raise distinct ethical concerns—dependency, commodification, and prolonged suspension of grief—that involve different relational dynamics from the bounded, ritualised encounters analysed here. Addressing such technologies would require a separate analytical frame and empirical base and therefore falls beyond the scope of this chapter. Future research should examine how these tools intersect with bereavement care under varying ethical and regulatory regimes.

These limits nonetheless point to critical directions for practice. Rather than mediated spectacle, future frameworks should prioritise co-created encounters that are explicitly bounded and culturally situated. The disruption introduced by AI lies in its capacity to simulate autonomous agency; yet, as Polanyi (1958) reminds us, machines lack commitment—the personal responsibility for truth and care. This gap highlights the need for ethical containment around immersive encounters, supported by trauma-informed guardrails that protect agency and enable withdrawal when needed (SAMHSA, 2014).

Crucially, such containment cannot be delivered by any single profession. The ethical and psychosocial vulnerabilities identified here indicate the need for multidisciplinary clinical engagement, involving palliative care clinicians, mental health professionals, social workers, and—where appropriate—ethics and governance expertise. Future research should therefore move beyond role assertions to examine how collaborative assessment, shared decision-making, and clinical governance can be integrated across preparation, encounter, and aftercare, and

how responsibilities are distributed within teams and institutions (Firth, Luff & Oliviere, 2005).

Taken together, these implications suggest that VR- and AI-enhanced bereavement care should remain provisional, relationally grounded, and ethically accountable. Across these multi-scalar dimensions, palliative care practitioners can ensure that virtual encounters support healing while also strengthening communities, shaping ethical institutions, and fostering transnational solidarity (Banks, 2020).

Conclusion: Carrying Memory, Returning to Life

This chapter shows that ethical use of VR and AI in bereavement care depends on their design as bounded, time-limited rituals. Virtual encounters can help the bereaved voice what was left unsaid and relocate bonds into an enduring inner presence (Klass, 1996; Stroebe & Schut, 1999). However, without safeguards, they risk becoming spectacles of indefinite suspension.

The proposed framework demonstrates that palliative care practitioners are not merely technicians of grief management but guardians of dignity working at the intersection of family care and civic justice (Banks, 2012; Walter, 2015). As argued through the Polanyian lens, this role requires *practical wisdom*—practical wisdom exercised in contexts where technical guidelines are insufficient. Bereavement practice is sustained by both *commitment*—the determination to preserve memory—and *calling*—the orientation of grief toward justice (Polanyi, 1958, 1966).

Ultimately, to carry memory is not to remain bound by the past but to weave loss into life in ways that sustain resilience. The true measure of a successful encounter lies not in the avatar's realism, but in the family's readiness to remove the headset and step away. Technology should serve not to conjure a 'digital ghost,' but to facilitate the internalisation of love and the recovery of everyday life. Held with humility and guided by tacit discernment, VR and AI can serve as a modest tool that honours the dead, steadies the living, and helps societies return to life—changed, but not diminished.

References

Alderson, P., & Morrow, V. (2020). *Ethics of research with children and young people: A practical handbook*. Sage.

Bailenson, J. (2018). *Experience on demand: What virtual reality is, how it works, and what it can do*. W. W. Norton & Company.

Banks, S. (2020). *Ethics and values in social work* (5th ed.). Bloomsbury Publishing Plc.

Berghs, M., Prick, A.-E. J. C., Vissers, C., & van Hooren, S. (2022). Drama therapy for children and adolescents with psychosocial problems: A systemic review on effects, means, therapeutic attitude, and supposed mechanisms of change. *Children*, *9*(9), 1358. https://doi.org/10.3390/children9091358

Boellstorff, T., Nardi, B., Pearce, C., & Taylor, T. L. (2012). *Ethnography and virtual worlds: A handbook of method*. Princeton University Press.

British Association of Social Workers. (2016). The role of social workers in palliative, end of life and bereavement care. https://new.basw.co.uk/policy-and-practice/resources/role-social-workers-palliative-end-life-and-bereavement-care

Brubaker, J. R., Hayes, G. R., & Dourish, P. (2013). Beyond the grave: Facebook as a site for the expansion of death and mourning. *The Information Society*, 29(3), 152–163. https://doi.org/10.1080/01972243.2013.777300

Buitelaar, J. C. (2017). Post-mortem privacy and informational self-determination. *Ethics and Information Technology*, *19*(2), 129–142. https://doi.org/10.1007/s10676-017-9421-9

Chochinov, H. M., Hack, T., Hassard, T., Kristjanson, L. J., McClement, S., & Harlos, M. (2005). Dignity therapy: A novel psychotherapeutic intervention for patients near the end of life. *Journal of Clinical Oncology*, *23*(24), 5520–5525. https://doi.org/10.1200/jco.2005.08.391

Firth, P., Luff, G., & Oliviere, D. (2005). *Loss, change and bereavement in palliative care*. Open University Press.

Floridi, L. (2013). *The ethics of information*. Oxford University Press.

Hall, S. (1997). *Representation: Cultural representation and signifying practices*. Sage.

Hine, C. (2000). *Virtual ethnography*. Sage Publications Ltd.

Hollanek, T., & Nowaczyk-Basińska, K. (2024). Griefbots, deadbots, postmortem avatars: On responsible applications of generative AI in the digital afterlife industry. *Philosophy & Technology*, 37, Article 63. https://doi.org/10.1007/s13347-024-00744-w

Holloway, M., & Taplin, S. (2013). Editorial. *British Journal of Social Work*, *43*(2), 203–215. https://doi.org/10.1093/bjsw/bct029

Hudson, R., & O'Connor, M. (2007). *Palliative care and aged care: A guide to practice*. Ausmed Publications.

Klass, D. (1996). *Continuing bonds: New understandings of grief*. Taylor & Francis.

Lee, B. (2020, February 8). Nayeon's mother turns blog private, citing fear of malicious comments ['악플 두렵다' 던 나연이 엄마, 블로그 비공개 전환한 이유]. Seoul Shinmun. https://www.seoul.co.kr/news/life/pop-culture/2020/02/08/20200208500023

Madary, M., & Metzinger, T. K. (2016). Real virtuality: A code of ethical conduct. Recommendations for good scientific practice and the consumers of VR-technology. *Frontiers in Robotics and AI*, 3. https://doi.org/10.3389/frobt.2016.00003

MBC (2020). [VR Human Documentary] Mother meets her deceased daughter through VR technology [Video]. YouTube. https://www.youtube.com/watch?v=uflTK8c4w0c

Munhwa Broadcasting Corporation (MBC). (2020). VR human documentary: I met you [Season 1]. iMBC. https://program.imbc.com/meetyou

Munhwa Broadcasting Corporation (MBC). (2021). VR human documentary: I met you [Season 2]. iMBC. https://program.imbc.com/meetyou

Munhwa Broadcasting Corporation (MBC). (2022). VR human documentary: I met you [Season 3]. iMBC. https://program.imbc.com/meetyou

Munhwa Broadcasting Corporation (MBC). (2023, 2024). VR human documentary: I met you [Season 4]. iMBC. https://program.imbc.com/meetyou

Neimeyer, R. A., & Thompson, B. E. (2014). Meaning making and the art of grief therapy. In B. E. Thompson & R. A. Neimeyer (Eds.), *Grief and the expressive arts: Practices for creating meaning* (pp. 3–13). Routledge.

Oh, S. (2025). How decisions are made for children and young people with life-limiting conditions: A critical ethnographic study [Doctoral dissertation, Durham University]. Durham e-Theses. https://etheses.dur.ac.uk/14232

Öhman, C., & Floridi, L. (2017). The political economy of death in the age of information: A critical approach to the digital afterlife industry. *Minds and Machines*, *27*(4), 639–662. https://doi.org/10.1007/s11023-017-9445-2

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. University of Chicago Press.

Polanyi, M. (1966). *The tacit dimension*. University of Chicago Press.
Polanyi, M., & Prosch, H. (1975). *Meaning*. University of Chicago Press.
Ramirez, E., & LaBarge, S. (2018). Real moral problems in the use of virtual reality. *Scholar Commons*. https://scholarcommons.scu.edu/phi/99/
Reith, M., & Payne, M. (2009). *Social work in end-of-life and palliative care*. Lyceum Books.
Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*(74). https://doi.org/10.3389/frobt.2016.00074
Stokes, P. (2012). Ghosts in the machine: Do the dead live on in Facebook? *Philosophy & Technology*, *25*(3), 363–379. https://doi.org/10.1007/s13347-011-0050-7
Stroebe, M., & Schut, H. (1999). The dual process model of coping with bereavement: Rationale and description. *Death Studies*, *23*(3), 197–224. https://doi.org/10.1080/074811899201046
Substance Abuse and Mental Health Services Administration. (2014). Trauma-informed care in behavioral health services (TIP 57). https://library.samhsa.gov/sites/default/files/sma14-4816.pdf
Turner, V. W. (1969). *The ritual process: Structure and anti-structure*. Aldine Transaction.
Walter, T. (2015). Communication media and the dead: From the Stone Age to Facebook. *Mortality*, *20*(3), 215–232. https://doi.org/10.1080/13576275.2014.993598
Worden, J. W. (2018). *Grief counseling and grief therapy: A handbook for the mental health practitioner* (5th ed.). Springer Publishing Company.

16 Bias in the Machine

How Artificial Intelligence Perpetuates Ableism in Society

Simona Aginskaitė and Eglė Šumskienė

Introduction

"We're sold this [technology] as a good thing, but it actually seems like a desperately sad world… partially genocidal, if you're a Deaf person and your culture will die out because of technological means" (E_04), reflects one expert's sharp critique of Artificial Intelligence's (AI) encroachment into disability domains. This tension between the promise of progress and the peril of cultural erasure captures the core paradox at the heart of AI's expansion into the disability field. The analysis presented in this chapter unpacks this contradiction by situating AI within two interrelated frameworks: technoableism and the supercrip narrative, both of which illuminate how technological solutions can reinforce, rather than dismantle, ableist norms.

The goal of this chapter is to examine how AI upholds ableist norms by pressuring people with disabilities to assimilate into dominant technological frameworks, often at the cost of their autonomy, cultural identity, or inclusion, and to critically analyze the implications of this dynamic for the future of disability justice.

The Evolution of Ableism into Technoableism

Ableism refers to discrimination and prejudice against people with disabilities, rooted in societal norms that prioritize able-bodiedness. Early conceptualizations of ableism positioned disability as a deviation from an ideal norm, reinforcing the idea that people with disabilities needed to be fixed or cured (Campbell, 2009). Ableist attitudes have emerged and evolved due to a combination of historical, cultural, scientific, and economic factors that have shaped how societies perceive disability. The key reasons and mechanisms through which ableism developed, include the historical construction of "normalcy" (Thomas & Sakellariou, 2018), the medicalization of disability (Haydon–Laurelut, 2015), economic structures and social perceptions (Clifford, 2020; Russel, 2019), cultural and media representations (Nario–Redmond 2019; Johanssen & Garrisi, 2020). Moreover, it intersected with other forms of oppression, such as colonialism and racialized ableism, gender inequality, and disability (Dirth & Adams, 2019). Technoableism builds upon ableist ideologies by offering new means to control or normalize people with disabilities, while the societal norms of prioritizing standardization over diversity are reinforced through technological

DOI: 10.4324/9781003627920-16
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

advancements. The medical model of disability has treated impairment as a personal deficiency, requiring cure, and contributed to technoableism by reinforcing the belief that technology could "normalize" bodies with disabilities rather than changing societal structures to accommodate them. The eugenics movement reinforced ableist hierarchies by promoting the idea that disability was undesirable and should be eliminated. Recent scholarly discourse suggests that technological advancements have led to the emergence of a new, technology-based eugenics movement (Smith, 1999; Marini & Stebnicki, 2019; Brown et al., 2019).

This example of technology-enabled shift in eugenics reflects a broader societal trend toward using technological interventions to manage and modify human traits, thereby perpetuating ableist ideologies under the guise of progress. The development of assistive technologies has been framed as a means of overcoming disability rather than embracing it. Bogart and Dunn (2019) note that AI-driven solutions frequently assume that people with disabilities should assimilate into able-bodied standards. This perspective aligns with the supercrip narrative, which is discussed below, where people with disabilities are celebrated for "overcoming" their impairments through technology, reinforcing the idea that disability is a burden to be minimized (Shew, 2023).

Cultural narratives, particularly the supercrip stereotype, play a crucial role in shaping public perceptions of disability. Media, as explored by Ellis and Kent (2011), portrays people with disabilities as heroic figures who overcome their impairments through exceptional effort or technological assistance, thereby reinforcing ableist expectations and overshadowing the systemic barriers they face. AI-driven solutions that highlight "miraculous" recoveries contribute to technoableism by suggesting that people with disability must either integrate seamlessly into technological systems or remain marginalized, further perpetuating the idea that technological integration equates to societal acceptance (Shew, 2023).

Since ableist attitudes permeate every aspect of the lives of people with disabilities, and technology is becoming increasingly ubiquitous across all facets of society, the consequences for individuals with disabilities are likely to be more perilous than promising. The use of these technologies, shaped by ableist assumptions, risks deepening existing inequalities and further marginalizing those who do not conform to normative standards of ability (Campbell, 2009). Rather than fostering genuine inclusion and accessibility, this trajectory may lead to the development of technological systems that reinforce ableist biases, solidifying societal barriers similar to those critiqued by Shakespeare (2013), who emphasizes social constructions contributing to disability exclusion. Technoableism is intricately linked with the supercrip narrative, both of which are rooted in ableist ideologies. Their interconnectedness and implications will be explored in greater detail in the subsequent section.

Technoableism and the Super-Crip narrative

Technoableism refers to the expectation that technology will "fix" disability, positioning AI as a tool that eliminates impairment rather than accommodating diverse

modes of existence. This perspective assumes that people with disabilities primarily seek technological solutions to their conditions, rather than broader societal accessibility and inclusion.

AI technologies reinforce technoableism by prioritizing innovations that attempt to normalize bodies with disability rather than adapting environments to diverse needs: exoskeletons for wheelchair users emphasize the ability to stand or walk; facial recognition systems fail to identify individuals with atypical facial features; and speech recognition software struggles with non-standard speech patterns. Marini and Stebnicki (2019) critically examine assistive technology, acknowledging its potential to enhance independence while also highlighting significant limitations. From a technoableism perspective, they discuss issues such as technology abandonment, emphasizing the role of user engagement, motivation, and environmental support in AI adoption.

Closely tied to technoableism is the supercrip narrative (Wolbring, 2024), which creates an implicit expectation that people with disabilities must achieve extraordinary feats to be valued by society. AI-driven assistive technologies reinforce this narrative by celebrating people with disabilities who successfully integrate into mainstream technological systems, rather than questioning whether those systems should be more inclusive: prosthetic limbs with AI-enhanced movement capabilities are marketed as "better than human" alternatives; AI-powered cognitive aids promote increased productivity in neurodivergent individuals; and automated hiring systems evaluate candidates with disabilities based on how well they conform to non-disabled communication and interaction norms.

Technoableism and the supercrip narrative are linked through their shared roots in ableist ideologies. The concept of technoableism involves the belief that technology can "fix" or eliminate disability, reflecting a broader ableist assumption that disability is inherently undesirable (Shew, 2023). Technoableism positions technology as a means to normalize disabled bodies and minds or offer a technological redemption (Shew, 2019). The supercrip narrative portrays individuals with disabilities as heroic figures who overcome their impairments through extraordinary effort or, increasingly often, through technological assistance (Wolbring, 2024). Cyborgs in media depictions are almost always the "apotheosis of human ability and appearance" (Shew & Earl, 2024, p. 6), further exemplifying how technological enhancement is framed as a means to transcend human limitations. Both technoableism and the supercrip narrative reinforce ableist norms by emphasizing the need for individuals with disabilities to conform to societal expectations and marginalize diverse experiences within the disability community. Technoableism does this by promoting technologies aimed at "overcoming" disability (Shew, 2023), while the supercrip narrative celebrates those who achieve remarkable accomplishments despite their impairments (Wolbring, 2024).

Methodology

The study aimed to explore how the AI shapes disability discourse through the perspectives of researchers, innovators, and disability activists with expertise linked to

the topic (hereafter referred to as *experts*). Semi-structured interviews were chosen as the data collection method, allowing for flexibility in the exploration of topics while maintaining consistency across interviews.

Data gathering for this research involved conducting semi-structured interviews with seven experts. This method is well suited to contexts where little is known about the subject, as it enables the collection of rich qualitative data and supports the exploration of how people construct meaning around new phenomena. A purposive sampling strategy was used to select interviewees with relevant expertise in AI and assistive technologies in disability. The interviews were carried out from November 2024 to March 2025. The interviews explored expert perspectives on the challenges of addressing disability in AI systems, including methods for involving individuals with disabilities in data contribution and technology development, and strategies to address participation barriers.

Interview recordings were transcribed verbatim and subsequently analyzed using thematic analysis. Seven key thematic categories emerged that are grounded in recurring themes and distinct analytical perspectives raised by the experts: Bias and Representation; Policy and Governance; Inclusive Design and Co-Creation; Techno-Ableism; Ethical Challenges; Economic and Practical Barriers; and Applications and Impact on Lived Experience.

In 2024, ethical approval was granted by the Research Ethics Compliance Committee at Vilnius University prior to the study (No. (1.13 E) 250000-KT-70), with all experts providing informed consent and ensuring confidentiality.

Data Analysis and Discussion

Bias and Representation

One of the strongest themes across the interviews is that current AI systems inherit or amplify biases against people with disabilities. The data and algorithms underpinning AI can reflect societal ableism: simply adding words like "visually impaired" or "ADHD" to otherwise neutral sentences caused the AI to give those sentences a negative sentiment score, even when no negativity was present in the original content:

> When ableist data enters AI systems, it acts as an even stronger categorising factor in those systems. I mean, if a text can have subtleties and nuances, then AI just categorises it into a simple good-bad category and the result is even more stereotyped (E01).

The mere presence of a disability identifier skewed the AI's judgment, revealing representational bias, which is not just a hypothetical concern: as demonstrated in a study previously conducted by one of the experts, it statistically manifested in the model's outputs, and it mirrors harmful stereotypes in society.

AI models misrepresent people with disabilities due to biased or incomplete training data. The robotics engineer observed that many AI models trained on older or non-representative data end up being "very ableistic" and even more prejudiced

than society (E_06). If the training data does not include diverse disabled experiences or is drawn from non-disabled perspectives, the resulting model will likely reflect those exclusions: if the system of an autonomous vehicle has not been trained on examples of people using wheelchairs, it may not detect them on the street, posing a direct safety risk (E_05). People with disabilities are statistical "outliers" in many datasets and thus get overlooked by pattern-needing AI models. It is practically impossible to capture the full diversity of disability in one universal dataset, and even well-intentioned universal design in AI will inevitably miss some people; moreover, optimizing a design for one disability may inadvertently create new barriers for another, leading to "access conflicts" (E_03).

AI bias can also compound harm within the disability community. Discussions about AI bias have focused more on gender and race, with disability bias only starting to get attention. Thus, the understanding of disability discrimination in AI is still lagging behind (E_02). This means fewer auditing tools and less awareness, which could allow biased systems to proliferate unchecked. Experts suggested the concept of *cumulative harm*—the idea that even minor biases in AI systems can add up over time and across platforms to significantly harm people with disabilities (E_02).

Policy and Governance

Interviewees discussed current legislative frameworks and whether they adequately address disability concerns. The European Union's AI Act takes a risk-based regulatory approach, categorizing AI systems by risk levels. However, what regulators consider a "low or medium risk" use of AI for the general population might actually pose high risks for persons with disabilities. For instance, a recruitment AI tool might be seen as low-risk overall, but could systematically filter out neurodivergent candidates (E_02).

Disability advocates lobbied to get accessibility mentioned in the AI Act and succeeded, but mainly with respect to high-risk systems. For lower-risk categories, explicit disability safeguards are still lacking (E_02). Companies in the tech industry feel that regulations "shouldn't limit business possibilities", seeing requirements for inclusion as costly obstacles (E_04). With so much AI development happening in private companies, and public funding for research declining, there is a risk that "it doesn't look that promising for people with disabilities" if profit-driven firms aren't mandated to prioritize accessibility (E_04).

While the EU AI Act is "meant to protect people", its broad definitions are so general that they might impede deploying beneficial innovations in real-world care settings:

> But these rules don't really cover the research stage — they apply at the point of deployment. The AI Act is written in very broad terms, for example, saying you shouldn't cause "significant harm" to a person. But what exactly counts as harm isn't spelled out clearly. That makes it hard to apply — both in research and in practice (E_05).

Certain helpful interventions (like nudging a patient for their own health) might be classed as high-risk or unethical by a strict reading of the rules. This highlights a governance challenge: how to craft rules that protect vulnerable users without unduly stifling potentially positive assistive technologies. The expert has consequently begun working with lawyers and ethicists to find the right balance, indicating that practitioners are actively engaging with governance frameworks to interpret and possibly refine them (E_05).

Inclusive Design and Co-Creation

All experts touched on the crucial importance of involving people with disabilities in the design and development of AI, not just as end-users, but as active co-creators. Many problems stem from a lack of voices of people with disabilities: there is a "mismatch between technology and the user's needs" because "the producer doesn't understand what the person's needs are, and the person with needs can't communicate them" (E_04). Experts advocate for approaches like Value Sensitive Design, where developers begin by deeply analyzing the daily practices and values in the context where the technology will be used, and only then proceed with stakeholder-driven design. In their research groups, experts promote doing this kind of groundwork "before you start even deciding", then iteratively involving relevant stakeholders (including end-users with disabilities) to define needs and spot ethical risks early. This "slow motion" approach is "not at all what's happening in business" and generates a tension between an inclusive, participatory design and industry timelines (unless there is a strong mandate to do so). Co-creation matters, otherwise "many AI inventions…don't make it past the prototyping stage," specifically in the context of disability tech (E_02). A lot of pilots ended up not being adopted because they weren't actually useful, or ignored practicalities like the availability of support services or funding streams to sustain the product.

Historically, the engineer's work pattern was "we see a problem, we decide what the problem is, we invent a solution, then hand it to the users and assume it's solved", which often failed (E_06). Now there is a shift toward including users at the design level, albeit still asymmetrically, with the designer in charge and the person with disability in a consulting role (E_06). Experts find value in participatory design and point to an emerging practice of having people with disabilities hired as part of the research team (E_06), thus lived experience is treated as a form of expertise equal to technical expertise. Moreover, such an aspiration aligns with the broader disability rights principle "Nothing about us without us".

Beyond mainstream inclusion, technologies designed by people with disabilities themselves better meet their needs and avoid outsider assumptions (E_03). Inclusion isn't only about inviting users to critique or contribute to nondisabled-led projects, but also about recognizing and supporting technologists and innovators with disabilities. Interviewees highlighted structural ways to support co-creation, such as a "testing infrastructure" where people with disabilities are paid to give feedback (E_02). Paying users, rather than relying on unpaid input, could enable more consistent, diverse participation and help make participatory design standard practice.

Techno-Ableism

Several experts referred to the tension between the perspective that sees technology primarily as a means to "fix" disability, and another that emphasizes accepting disability as a valuable diversity, as techno-ableism—when technology is driven by ableist assumptions that people with disability should strive to become as "normal" as possible.

When society imagines what is "good" for people with disabilities, it assumes all of them wish to be "fixed". This mindset can drive technology development in directions that undermine disability identity or pride: "sometimes when people imagine what's good for people with disabilities, they imagine all of them wish to be non-disabled" (E_03). A classic example is the enthusiasm around high-tech exoskeletons intended to help wheelchair users walk: "tech people think [the exoskeleton] is going to be the primary device… and people were like, 'well, my wheelchair is pretty great. Why does it always have to be one or the other?'" (E_03). Ableist bias in tech can frame a „normal" life as one that avoids using mobility aids, whereas persons with disabilities perceive them as enabling tools. Experts agreed that assuming "walking = better" or "hearing = better" without listening to those who live with a disability is a flawed starting point.

The debate over cochlear implants for Deaf individuals was repeatedly cited as emblematic of this clash. Experts discussed the "cochlear implant dilemma": many parents of Deaf children and medical professionals champion implants as a miraculous cure for deafness; whereas members of Deaf communities resist the idea that deafness needs curing, seeing it as a linguistic minority status. The expert empathized with parents' desire to provide opportunities for their child, but also conveyed the Deaf community's fear that the widespread adoption of cochlear implants could erode Deaf culture. This debate raises an ethical tension between normative individualism (doing what is best for an individual child, giving them hearing) and collective cultural rights (preserving a minority language).

Once a technology exists, even as an option, social pressure can arise: those who decline it may be stigmatized as "choosing to stay disabled." It is vital that tech does not become an excuse to withdraw other accommodations or to deem people who don't use the latest device as less worthy. In essence, society should continue to become more accessible regardless of new medical or assistive "cures," and this perspective argues that things like the exclusion of people with disabilities are social issues—you cannot simply "techno-fix" prejudice or inaccessibility by altering a person's body or abilities.

Techno-ableist narratives can even harm people with disabilities economically and socially: when technology is touted as a "solution", it can lead to scenarios where individuals with disabilities feel compelled to adopt expensive or invasive tech just to be seen as acceptable; otherwise, they risk further marginalization. Experts warned of a potential future where technologically enhanced individuals set a new bar for ability, and those who remain unenhanced are seen as a disadvantaged class—a dystopia of stratified ability (E_07) and put it pointedly: "You can't technologize yourself out of being bullied" or discriminated against; if the root prejudice isn't addressed, the target simply shifts (E_07).

Ethical Challenges

The computer scientist discussed how their team uses techniques like nudging or even small deceptions in social robots to encourage patients to follow medical advice. While they noted that such strategies can be effectively used by therapists or parents in certain situations, they are acutely aware that when AI does this, it "of course has some ethical and legal concerns" (E_05). The ethical dilemma is that the robot is deliberately steering the person's behaviour by exploiting psychological levers, potentially "tailoring the behaviour…towards the vulnerabilities" of the patient (E_05). Even if the intent is benevolent, it raises questions: How much deception is it okay for a care robot to use? At what point does helpful persuasion become a violation of a person's informed consent or dignity? Similarly, the philosopher offered a thought experiment about AI "mental extenders" for people with cognitive disabilities or ADHD and imagined AI tools that could, say, subtly reshape someone's preferences or assist in decision-making. Any technology that intervenes in "your preference formation, your belief formation…always carries the risk of having a causal impact on what you really desire or prefer, and that's seriously problematic for autonomy" (E_04).

Another ethical concern is privacy and surveillance, which disproportionately affect people with disabilities through social service checks or medical monitoring and may be intensified by AI. One expert's friend with a computerized prosthetic knee counts steps and is anxious that if she doesn't log "enough" steps, an insurer might deem her unworthy of the device, but if she exceeds expectations, they might also take it away, assuming she is "too functional" (E_03). As the expert put it, "the sort of counting and scrutiny of our bodies" via data is worrisome: every action on a smart device creates logs, and "that can be weaponized against me" (E_03). People with disabilities might get flagged by automated systems simply for being different. AI systems deployed for security can disproportionately harm individuals with disabilities by misclassifying disability-related behaviour as suspicious. High-tech solutions should not replace humane approaches: if someone doesn't accept a person with disability for who they are, giving that person with disability a fancy technology won't change the prejudice, and might even add new stigma (E_07).

Safety is another concern, as many AI systems are essentially "black boxes"—their internal decision logic is opaque; it's hard to predict or explain their behaviour, which is a safety issue for users. We "never know the exact decision-making" of these models (E_01): if an AI healthcare tool denies a disability benefit or a social service, both the user and the officials might not know why it made that call, making it difficult to fix errors (E_01, E_02).

When it comes to robots and AI caretakers, emotional safety and well-being also arise. Social robots can provide companionship to lonely individuals or repetitive patience to someone with dementia, but some people may become too attached to robots or prefer them to humans, which would ultimately deepen their isolation. People who "suffer from a bit of disconnection" might confuse robot companions for real, reciprocal relationships (E_05): an elderly person might emotionally bond with a robot pet and then experience distress because that bond is not "real". If

such a robot were later taken away, the person could be worse off emotionally than if they had never had it. Ethically, this raises questions about designing robots that deliberately encourage emotional attachment. The robotics expert echoed this concern, noting that "sometimes, particularly with people who suffer from loneliness, the robot becomes a main part of their life," which in some cases can exacerbate psychological issues (E_06). On the other hand, experts recognized that for some individuals, some interaction (even with a machine) is better than none.

Economic and Practical Barriers

Even the most innovative AI solution can be rendered moot if it is not accessible or affordable to those who need it. A significant theme in the interviews was the set of economic and practical challenges that determine whether AI technologies actually benefit people with disabilities on a broad scale. Experts pointed out that high costs, market incentives misaligned with inclusion, and a lack of support infrastructure can all impede the equitable distribution of beneficial tech.

Many advanced technologies are extremely expensive. The vast majority of people with disabilities would "never be able to afford" these innovations (E_07). Historically, approaches like Community-Based Rehabilitation arose precisely because giving everyone personalized high-tech devices was not feasible in poorer regions; instead, communities focused on low-cost environmental modifications and peer support (E_07). This perspective suggests that while AI devices capture headlines, simpler, cheaper solutions, like making schools and workplaces accessible, might yield more immediate benefits to more people. There is a risk that a focus on high-tech cures diverts attention from these more accessible solutions. The expert also mentioned research documenting that many people with disabilities do not want certain assistive technologies either because the devices don't meet their needs or because they prefer managing without them rather than using something cumbersome or stigmatizing (E_07).

Another economic barrier is that even when good assistive technologies exist, getting them to individuals requires strong support systems: a fancy assistive device might not get adopted in a country if there is no policy for insurance or no training program for users, making it effectively infeasible (E_02). Even if a future technology could replace wheelchairs, if it's not universally affordable and serviced, society cannot rely on it; we must continue to invest in basic accessibility. Moreover, technology can fail or break (E_06).

Experts painted a potential scenario where advanced enhancements create a new class divide: wealthier people with disabilities become more competitive in workplaces, while those who cannot afford enhancements fall further behind (E_07). This could worsen internal disparities within the disability community. Without interventions, we risk a hierarchy where, for example, one amputee with a bionic arm might outperform another amputee with a basic prosthesis, and society might start expecting the latter to "upgrade or step aside" (E_07).

Experts argue that an equitable approach is to invest in public goods and universal design that benefit everyone, with or without disability, rather than

super-specialized tech for a few. For example, making AI voice assistants better at understanding slurred speech or different communication modes might help not only people with speech disabilities but also the general population. Ensuring websites and apps meet accessibility guidelines benefits a broad user base. These kinds of improvements have relatively low marginal cost and high social benefit, but they may not happen without policy enforcement.

Lastly, the experts highlight the importance of holistic thinking, and lack thereof and an "overemphasis on just the technical perspectives" lead to failures (E_02). Holistic planning includes training users, supporting low-cost or open-source AI, and valuing simple, scalable community solutions over expensive high-tech options (E_03).

Applications and Impact on Lived Experience

Experts highlighted several promising use cases, such as a social robot deployed in schools to conduct an initial screening for eating disorders among teens: the robot engages with students to gather health data, working from a script developed with psychologists (E_05). This sort of application shows AI's potential to augment healthcare: it can perform routine yet sensitive tasks in a way that might even feel less stigmatizing to some students than talking directly to an adult. Robots can also be used to ease patient anxiety during medical procedures: a robot distracts children while they receive a shot (E_05). A social robot is used to encourage adherence in cardiac rehabilitation patients. Such rehab requires patients to come exercise regularly for months, and dropout rates are high. The robot acts as a coach to motivate patients to keep attending sessions, and the robot's social presence is the intervention. The result, ideally, is improved health outcomes for patients because they stick with their therapy (E_06). Using AI as avatars or extensions of a person can be transformative for social inclusion, as technology enables participation that would have been impossible before. The expert sees such robots that combine physical assistance and social interaction as the future of assistive tech (E_06).

However, the interviews also discuss limitations and unintended consequences of these applications. One recurring concern is the risk of dehumanization or reduced human contact. Some elderly individuals or those with a disability might end up with most of their interaction being with a machine if a care home over-relies on robotic caregivers (E_05). If a social robot greets you, feeds you, and reminds you to take medicine, but no one else comes to spend time with you, loneliness could persist or worsen. Experts mentioned cases where people might even prefer the robot because it doesn't judge or get impatient, which might lead them to withdraw from human relationships (E_05). Ethicists would call this the "substitution effect": the robot is meant to supplement, but it might substitute for human care due to cost-cutting or convenience, and that's ethically fraught.

Nurses might shift into more supervisory or technical roles, managing AI outputs, and focus more on interpersonal aspects of care, or it could lead to deskilling in some areas. Big tech companies likely have the know-how to make things accessible, but they sometimes balk due to expense—integrating AI could either lighten caregivers' workload or be seen as a way to cut staffing (E_02, E_05).

There are also cultural impacts: as certain assistive technologies become more advanced and even fashionable, it might change social perceptions of disability. An advanced prosthetic arm could even become a kind of status symbol or at least not something to be ashamed of, perhaps reducing stigma (E_03). However, this verge into "transhumanism" introduces new questions: if prosthetics surpass typical human abilities, are we creating pressure for people with disabilities to go beyond "normal" to be accepted (E_03, E_07)?

Many experts celebrated how AI and tech, when done right, can increase independence for people with disabilities. Technologies like exoskeletons or brain interfaces are ethically sound if they indeed give an individual more autonomy, and it's "what they want" (E_04). For permanent disabilities, the aim of the AI is to "improve their lives with the disability, not to eradicate [it]", by providing tools that grant more freedom (E_05).

Conclusion

This study confirms and advances the concept of technoableism as a critical framework for understanding how AI is reshaping the disability field. Technoableism refers to the ideological expectation that technology should "fix" disability, often sidelining the broader social and structural changes needed to foster real inclusion. Through empirical data drawn from expert interviews, this research demonstrates how technoableism operates not only discursively but materially, throughout design processes, policy frameworks, and the everyday realities of people with disabilities interacting (or refusing to interact) with AI.

Technoableist assumptions manifest in all thematic categories emerging from the data and reveal a recurring pattern: AI systems are frequently designed to make people with disabilities more "acceptable" to existing norms, rather than making society more accommodating to difference. Likewise, the prioritization of "supercrip" narratives, where individuals with disabilities are celebrated only when they overcome impairments via high-tech solutions (Wolbring, 2024), continues to marginalize those who resist or cannot access such pathways.

The findings outlined in this chapter contribute new layers to existing scholarship. First, the concept of cumulative technoableist harm emerges as a powerful analytical tool, showing how repeated minor misrepresentations in AI systems compound into significant exclusion. This insight extends the critique of technoableism into governance, revealing a lack of disability-centred regulation that challenges the market-driven logic framing inclusion as a cost rather than a design priority. Second, the study highlights how technoableism is sustained through economic and regulatory structures that disincentivize inclusive design: co-creation with users who have disabilities is often undervalued, unpaid, or tokenistic, despite its proven potential to reduce technology abandonment and increase ethical alignment.

This study provides both confirmation and expansion of the technoableism framework. It shows how AI-driven disability innovations are shaped by long-standing cultural, economic, and political forces, and suggests that unless these forces are actively confronted, technological progress will continue to deepen

existing inequalities. The future of AI in disability must not be measured by how efficiently it normalizes difference, but by how radically it reimagines inclusion. Scholars such as Clare (2017) and Shew (2023) argue that genuine accessibility is not achieved through technological assimilation but through collective restructuring of the social order to embrace human diversity, or, as one of the experts put it: "if we're using technologies to solve social problems, it's never a good idea… social problems need social remedies" (E_07).

Implications

The data suggests pathways toward more emancipatory possibilities. Meaningful involvement of people with disabilities offers benefits beyond addressing specific functional needs. It elevates lived experience to the level of high-value expertise, enabling more inclusive and contextually grounded technological design. Experts emphasized the value of technologies designed by people with disabilities, not merely for them, challenging the hierarchical model of innovation and promoting the principle of "Nothing about us without us": inclusive, disability-led design is more likely to produce technologies that genuinely improve lives rather than reinscribe ableist norms.

Data shows how the role of designers is shifting—from creating discrete tools or functionalities to shaping vehicles for participation, inclusion, and attitudinal change. However, this transformative potential also carries risks: designers' creations can function as delayed-action mechanisms, with uncontrollable triggers and unpredictable effects. Design practices reshape social structures, influence societal cohesion, increase social distance, and alter the roles and responsibilities of traditional professions. Moreover, while they may help to challenge certain forms of stigma, they can simultaneously produce new and unforeseen forms of discrimination. Designers, developers, and innovators themselves possess a form of lived experience within the innovation ecosystem—embodied knowledge of creation processes, technical constraints, and emerging possibilities. When this insider perspective intersects with disability expertise, it exposes how innovation cultures operate and how design choices carry embedded social meanings, thereby shaping the trajectory of technological development. This combined expertise could offer unique insights into the (mis)alignment between innovation practices and regulatory intentions, making it a valuable resource for informing policy and legislative processes in ways that reflect both technical realities and social implications.

References

Bogart, K. R., & Dunn, D. S. (2019). Ableism special issue introduction. *Journal of Social Issues, 75*(3), 650–664. https://doi.org/10.1111/josi.12354

Brown, I., Brown, R. I., & Schippers, A. (2019). A quality of life perspective on the new eugenics. *Journal of Policy and Practice in Intellectual Disabilities, 16*(2), 121–126. https://doi.org/10.1111/jppi.12299

Campbell, F. K. (2009). *Contours of ableism: The production of disability and abledness*. Palgrave Macmillan.

Clare, E. (2017) *Brilliant imperfection: Grappling with cure*. Duke University Press.
Clifford, E. (2020). *The war on disabled people: Capitalism, welfare and the making of a human catastrophe*. Bloomsbury Publishing.
Dirth, T. P., & Adams, G. A. (2019). Decolonial theory and disability studies: On the modernity/coloniality of ability. *Journal of Social and Political Psychology, 7*(1), 260–289.
Ellis, K., & Kent, M. (2011). *Disability and new media*. Routledge.
Haydon-Laurelut, M. A. (2015). Disability: Beyond individualisation, psychologisation and medicalisation. *Metalogos*, *27*, 1–15. https://pure.port.ac.uk/ws/portalfiles/portal/2957423/Disability_Beyond_individualization.pdf
Johanssen, J., & Garrisi, D. (Eds.). (2020). *Disability, media, and representations: Other bodies*. Taylor & Francis.
Marini, I., & Stebnicki, M. A. (2019). *The psychological and social impact of illness and disability* (7th ed.). Springer Publishing Company.
Nario-Redmond, M. R. (2019). *Ableism: The causes and consequences of disability prejudice*. Wiley Blackwell.
Russell, M. (2019). *Capitalism and disability: Selected writings by Marta Russell*. Haymarket Books.
Shakespeare, T. (2013). *Disability rights and wrongs revisited*. Routledge.
Shew, A. (2019, April 23). Stop depicting technology as redeeming disabled people. *Nursing Clio*. https://nursingclio.org/2019/04/23/stop-depicting-technology-as-redeeming-disabled-people/
Shew, A. (2023). *Against technoableism: Rethinking who needs improvement*. W. W. Norton & Company.
Shew, A., & Earle, J. (2024). Cyborg-technology relations. *Journal of Human-Technology Relations*, *2*(1). https://doi.org/10.59490/jhtr.2024.2.7073
Smith, J. D. (1999). Thoughts on the changing meaning of disability: New eugenics or new wholeness? *Remedial and Special Education, 20*(3), 131–133. https://doi.org/10.1177/074193259902000302
Thomas, G. M., & Sakellariou, D. (2018). Introduction: Disability, normalcy, and the everyday. In G. Thomas and D. Sakiellariou (Eds.). *Disability, normalcy, and the everyday* (pp. 3–16). Routledge.
Wolbring, G. (2024). From supercrip to techno-supercrip. *Current Issues in Sport Science (CISS)*. Advance online publication.

17 Artificial Intelligence between the Lighthouse and the Storm

Potentials and Risks for Social Work[1]

Renata Nunes, Melisa Campana Alabarce, David Alonso González, and Andrés Arias Astray

Introduction

We are currently immersed in a profound technological transformation that is reshaping the contours of contemporary life. Among the most influential drivers of this change is artificial intelligence (AI), which is emerging not merely as a technical innovation but as a catalyst of structural and systemic shifts across all domains of human activity (Brandao, 2025). Its growing presence across sectors has generated both enthusiasm and concern, particularly in the human and social sciences, where it challenges fundamental notions of agency, ethics, and justice (Garkisch & Goldkind, 2025; Reamer, 2023). Within this context, social work faces a pivotal moment in which the promise of technological innovation must be carefully weighed against the risks of epistemological, methodological, and ethical displacement. AI technologies can transform education, research, and practice by automating tasks, enabling predictive analytics, personalising learning, and optimising decision-making. Yet, these possibilities are accompanied by profound challenges such as algorithmic bias, lack of empathy, data privacy concerns, unequal access to technology, and the erosion of human-centred judgement. Taken together, these risks reveal that the integration of AI into social work goes far beyond technical or organisational considerations, representing an ethical, political, and epistemological transformation that demands critical reflection, continuous human oversight, and a renewed commitment to social justice and professional responsibility.

Although AI offers new possibilities for efficiency, personalisation, and innovation, its rapid and often uncritical adoption also introduces significant risks and dilemmas for social work. Overreliance on algorithmic systems can erode professional autonomy and weaken the relational and empathetic dimensions that define the discipline. Data-driven tools may reproduce social and cultural biases embedded in training datasets, leading to discriminatory or decontextualised decisions. Unequal access to technological resources can deepen existing inequalities and create new forms of exclusion among both professionals and service users. Furthermore, the opacity of algorithmic processes raises serious concerns regarding accountability, informed consent, and the protection of vulnerable populations.

Against this backdrop, this chapter explores how AI may reconfigure the field of social work through a series of hypothetical case studies that simulate plausible

DOI: 10.4324/9781003627920-17
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

futures in four key domains: education and training, research, intervention, and organisational decision-making. These scenarios are not descriptions of existing practices, but speculative explorations designed to illuminate both the opportunities and the risks associated with AI's integration into the profession. Drawing on current debates on ethical governance and accountability, including the regulatory framework established by the EU Artificial Intelligence Act (Regulation 2024/1689), the chapter seeks to foster critical reflection on how social work can adapt and contribute to shaping human-centred, socially just uses of AI. Through these simulated settings, we aim to support the profession's ethical vigilance and epistemic resilience, while encouraging practitioners and educators to envision how future social workers might develop the competencies and reflexive capacities required to engage responsibly with AI-driven systems.

The chapter is organised into four main sections, each addressing a key domain of social work affected by the integration of AI. The first section, Pedagogical Implications of AI in Social Work Education, examines how AI-based platforms and simulations are transforming teaching and learning processes, highlighting both the opportunities for personalised education and the risks of overreliance and digital inequality. The second section, Research Dimensions of AI in social work, explores how algorithmic tools and big data analytics are reshaping research methodologies, raising epistemological questions about knowledge production and interpretive diversity. The third section, AI Implications for Social Work Intervention, focuses on professional practice, analysing how predictive systems and automated procedures influence case management, ethical decision-making, and professional autonomy. The fourth section, AI Implications for Social Organisations and Decision-Making Processes, investigates the governance and accountability dilemmas that emerge when algorithmic systems are implemented in institutional contexts. Each section develops a hypothetical scenario to illustrate concrete opportunities, risks, and ethical tensions. The chapter concludes by synthesising the main insights from these scenarios and reflecting on the political, institutional, and ethical frameworks necessary to ensure that AI contributes to human dignity, inclusion, and social justice in social work.

Pedagogical Implications of AI in Social Work Education

AI-driven platforms, from intelligent tutoring systems and adaptive algorithms to immersive simulations, are reshaping social work education. They promise broader access, personalised learning pathways, and realistic practice for high-stakes scenarios (Wang et al., 2023). At the same time, they raise concerns about commodifying learning, eroding critical interaction, and widening the digital divide. Preserving the discipline's humanistic ethos, therefore, requires models in which technological efficiency remains subordinate to ethical deliberation, relational engagement, and social justice (Nayir, Sari & Bozkurt, 2024; Vesna et al., 2025). To interrogate the contemporary challenges and dilemmas facing higher education, we propose hypothetical scenarios through a reflective and forward-looking methodology, giving concrete shape to practical situations that confront us in an era marked by the advance of AI.

Hypothetical Scenario: Teaching Experience

At a European public university, a social work lecturer, trained before AI became commonplace, implements the institution's new tools with measured caution. Using an intelligent planner, she reorganises the syllabus, updates readings, and adapts materials to diverse learning profiles. In her Professional Ethics module, a simulation generator produces short videos with digital avatars, encouraging debate and ethical reflection before students face real cases. Natural language chatbots support practice in empathy and active listening; an adaptive platform tracks individual progress, while an AI-assisted marker accelerates routine assessment and frees the lecturer for qualitative feedback. Each unit concludes with an interactive quiz that provides immediate results, sustaining engagement without compromising rigour.

Hypothetical Scenario: Student Experience

Juan, a third-semester student, illustrates the tension between innovation and caution. He consults the adaptive study plan on his phone (summarised readings, support videos, and a timetable matched to his pace), yet tiredness from night work persuades him to postpone the full readings, weakening sustained engagement with primary sources. In class, the lecturer screens a simulation of conflict between welfare policy and the right to housing; realism enlivens discussion but reveals algorithmic assumptions requiring critical scrutiny. Students present intervention proposals to a client chatbot on university tablets; device failures expose infrastructural limits and inequalities of access.

During the debate, the lecturer (Gabriel) introduces conceptual frameworks, prompts critical reflection, and highlights ethical dilemmas, guiding Juan to refine his argument in response to the chatbot's challenges. That evening, the adaptive dashboard identifies gaps in his risk-assessment skills and recommends a 15-minute micro-course and short test, which he completes before uploading an essay. Automated feedback on citation and coherence arrives within minutes, yet he waits for qualitative comments to capture nuances beyond the algorithm's scope. The unit ends with a gamified quiz; motivated but unsatisfied with its treatment of structural inequality, Gabriel prepares questions for the next face-to-face session.

Incremental Potentials of AI

Evidence from Papaspyridis and La Greca (2023) suggests that AI can enhance personalisation, sustain engagement, and release lecturers for more complex tasks. In this scenario, an intelligent planner reduces administrative load; concise simulations provide immediate discussion triggers; chatbots enable low-stakes communication practice; an adaptive dashboard helps students target urgent knowledge gaps; and automated marking accelerates technical feedback. These gains are incremental rather than transformative and improve efficiency and focus, but they still require sustained human oversight and critical interrogation of underlying assumptions and first-order questions.

In the context of social work education, Toli and Manasa (2024) argue that, in addition to personalising instruction, AI technologies can also support multiple dimensions of student learning. These include enabling practice with virtual clients, assisting in case analysis by identifying critical patterns, enhancing counselling abilities via chatbots, and delivering tailored learning pathways through intelligent tutoring systems. Furthermore, immersive virtual reality simulations offer experiential learning opportunities through realistic client interactions. At the same time, AI also facilitates several tasks of scientific work, such as bibliographic searches, text summarisation, writing, and citation.

In the context of integrating AI into social work education and intervention, further research is needed to identify the genuinely transformative gains—beyond the incremental benefits already observed—and the pedagogical and institutional conditions under which they might materialise. This research agenda should clarify how AI could reconfigure students' modes of thinking and reasoning; which instructional and assessment designs foster the generation of unforeseen options rather than the optimisation of what already exists; which metrics capture deep changes (transfer of learning, agency, abductive reasoning) beyond speed and technical accuracy; and how to balance algorithmic personalisation with exposure to diversity and dissent. It is also important to examine the accompanying epistemic risks—homogenisation of knowledge and solutions, biases in the generation of alternatives, and overfitting to historical patterns—and to specify the role of academic staff as mediators and critical curators, including the integration of situated epistemologies and community knowledge into AI workflows. Ultimately, it is necessary to determine whether AI can enable substantive transformations in education and practice without eroding the principles of social justice that orient the discipline.

Emerging Pedagogical Dilemmas

These benefits are inseparable from serious dilemmas. When critical reading, reflective writing, and ethical deliberation are delegated to machines, heavy reliance on algorithmic assistants can compromise students' intellectual autonomy. Moreover, if students turn to AI from the outset of their education, there is a risk that they will not effectively learn foundational skills—for example, synthesising readings, outlining an intervention plan, and producing notes and case records. Such tasks become routine for experienced practitioners; yet, without deliberate practice, students remain dependent on the tool and fail to consolidate these abilities. The result is a superficial, technology-contingent competence that may weaken professional judgement precisely when nuanced human reasoning is most needed. Junqueira notes that the spread of generative AI unsettles traditional understandings of authorship, creativity, and power, alters human–machine relations, and affects cognitive processes such as memory, reasoning, and intelligence.

Biases related to gender, class, and culture can permeate AI-generated materials, reproducing stereotypes the profession seeks to dismantle. The use of sensitive data in simulations threatens confidentiality, while algorithmic standardisation

flattens diverse learning trajectories (Baker & Hawn, 2022; Reamer, 2023). Community knowledge and Southern epistemologies remain largely absent from datasets trained on Northern sources, for example (Arora et al., 2023). Continuous monitoring may extend institutional surveillance without informed consent, and paywalled tools threaten to create new fault lines between digitally privileged and disadvantaged students. Finally, rapid automated feedback can displace the deep pedagogical dialogue essential to social work.

Robust safeguards are therefore required: systematic human oversight of algorithmic outputs, transparent audits of data and models, stringent privacy and consent policies, comprehensive AI literacy for staff and students, and equitable access to technology. With such measures in place, AI can strengthen, rather than undermine, social work's commitments to care, ethical deliberation, and social justice.

Research Dimensions of AI in Social Work

AI is increasingly used in research methodologies through machine learning, natural language processing, and big data analytics. These tools enable researchers to process massive data sets and detect patterns previously unnoticed. This section analyses how AI can empower evidence-based practice while also questioning the epistemological implications of relying on algorithmic knowledge production. It critically examines risks such as data quality, the potential replication of societal biases in training data, and the reduced role of qualitative, narrative, and participatory research methods. It also acknowledges that the very competencies required for conducting research, along with the imaginative capacities of researchers, may themselves come under scrutiny. Furthermore, unpacking the implications embedded in the analytical algorithms' underlying assumptions presents its own set of difficulties and added risks.

Hypothetical Scenario 2: Social Work Research

At an interdisciplinary social research institute affiliated with a public university, a social work research team launches a project to analyse housing insecurity among precarious migrant communities. With funding aimed at innovation and digital transformation, the team incorporates AI into their research design. Natural language processing (NLP) tools are used to analyse thousands of policy reports, online forums, and administrative case files, surfacing discursive trends and institutional blind spots. A predictive machine learning model trained on publicly available housing data identifies patterns of eviction risk across municipalities.

An automated transcription platform facilitates the processing of qualitative interviews, while visual analytics dashboards aid researchers preparing for participatory workshops with community stakeholders. The project's lead researcher, grounded in feminist and decolonial methodologies, balances algorithmic insights with local narratives and interpretive depth, emphasising human judgement and epistemic pluralism. Weekly team meetings include audits of AI-generated outputs to ensure transparency and alignment with social work's ethical principles.

Hypothetical Scenario: Researcher Experience

Carla, a doctoral researcher in the team, is intrigued by the speed at which NLP engines identify latent categories in policy discourse. Yet, as she explores the outputs, she notices that resilience narratives of migrant women are being misclassified as passivity, an issue later traced back to the training dataset's reliance on Eurocentric emotion ontologies, confirming the warnings outlined by Černý (2024). When presenting findings at a policy dialogue, Carla overlays algorithmic forecasts with counter-stories from affected communities to resist data-driven essentialism.

During qualitative coding, she finds the automated system fails to register silences, metaphors, or ambivalent expressions that carry crucial meaning for trauma-informed inquiry. She convenes a reflexive debrief with colleagues to revise the model's classification logic. To safeguard epistemic integrity, she organises a series of co-analysis sessions with migrant participants, ensuring research outcomes remain grounded in lived experience.

Incremental Potentials of AI

Recent work in social science confirms that AI can streamline research workflows, particularly in preprocessing and exploratory analysis. According to Dritsas and Trigka (2025), machine learning and NLP tools assist in processing large, unstructured datasets, improving hypothesis development and knowledge mapping. Carla's experience confirms that AI can support, though not replace, critical interpretation when guided by interdisciplinary ethics.

In a related review, Peponakis, Kapidakis and Doerr (2024) describe how computational ethnography and social anthropology benefit from AI-enhanced scale, provided interpretive sovereignty is preserved. These tools amplify reach, enable rapid feedback loops, and improve visibility of findings through dashboards and dynamic interfaces, but their value lies in augmenting rather than automating human reasoning.

Emerging Epistemological Dilemmas

AI-enhanced research presents critical methodological and ethical challenges. When knowledge production is outsourced to models, there is a risk of privileging computational legibility over contextual meaning. Riley et al. (2025) caution against the overuse of automated synthesis, warning that deep epistemic engagement can be diluted by pressure to produce rapid, scalable insights. Automated systems often miss subtle cues essential in trauma, cultural nuance, or resistance narratives.

As Bustio-Martínez and Herrera-Semenets (2024) explain, generative models trained on mainstream datasets risk replicating colonial or class-based epistemologies. Proprietary tools and paywalls further restrict epistemic justice, excluding under-resourced scholars and communities. Therefore, AI development must consider plural worldviews, especially in social fields driven by care and justice.

Consequently, robust safeguards are essential: epistemologically diverse datasets, model interpretability, participatory validation of outputs, interdisciplinary supervision, and AI literacy for researchers and participants. With such measures, AI can enhance the critical, relational, and justice-oriented commitments of social work research.

Furthermore, the increasing reliance on AI tools in research raises critical questions about the evolving nature of research competencies and the role of investigative imagination. When complex analytic procedures are delegated to algorithms, researchers may risk narrowing their capacity for conceptual exploration and critical interpretation. Moreover, unpacking the implications embedded within the underlying assumptions of algorithmic models is a demanding task, laden with epistemological, cultural, and ethical risks. As such, cultivating the ability to interrogate these foundations critically becomes as essential as mastering the technical tools themselves.

AI Implications for Social Work Intervention

Here, the focus shifts to the practical application of AI in daily intervention contexts, like case triage systems, automated referrals, and predictive risk tools. While such tools may enhance responsiveness, reduce workloads, and optimise resource allocation, they risk marginalising the human relationship at the heart of social work. This section critically examines the trade-off between automation and empathy, and questions whether AI can or should replicate professional judgement, especially in situations involving complex trauma, culture-specific contexts, and ethical dilemmas.

Hypothetical Scenario

In a medium-sized European city, following a major digitalisation process, the social services administration decides to introduce AI tools into the primary social-care network. The measure responds to months-long waiting lists; social work teams overburdened with assessments, reports, and unforeseen emergencies; and a high volume of bureaucratic tasks. The new platform is implemented to streamline assessments, plan interventions, and automate record-keeping. Within a few months, it becomes central to day-to-day operations. The initial appraisal is positive: the AI generates draft assessments within seconds, proposes intervention plans on the basis of historical case patterns, and automatically produces minutes and case notes, thereby freeing time for direct client work.

Hypothetical Scenario: Professional Experience of the Social Worker

The following scenario, while fictional, presents a plausible example of how AI tools, already available in areas such as transcription, automated summarisation, and case management, could be integrated into everyday social work practice.

Vanina, a social worker with ten years' experience, integrates these tools prudently. During interviews, she dictates notes or activates automatic transcription; at the end, the platform proposes a preliminary plan with objectives, responsible parties, and temporal milestones. Vanina reviews and contextualises each section, adjusts the language and priorities, and ensures that the document reflects her professional judgement, rendering it ready within minutes. In support groups (families undergoing rehousing, carers of older adults), she activates automatic minute-taking: the system transcribes, identifies agreements and tasks, and generates a structured summary with next steps and dates, which is shared immediately with the group and the interdisciplinary team, thereby improving continuity of care.

After a year, however, signs of dependence on the platform emerge. Some of Vanina's colleagues note that they tend to place excessive trust in the algorithmic proposals and that their critical review has weakened. The standardisation of assessments and plans homogenises readings of distinct situations and diminishes the team's analytical capacity; it even becomes more difficult to recall the particularities of each family and personal history. An emblematic case illustrates the problem: for a migrant family experiencing domestic violence, the tool recommended inappropriate measures because it failed to take adequate account of racialisation and the community context. The inconsistency went unnoticed until the conflict escalated.

In parallel, waiting lists decrease, but so does the number of professionals. The administration argues that, owing to the efficiency of AI, maintaining previous staffing levels is no longer justified, and it promotes redeployment to tasks of system oversight or technical support for the platform, leaving grey areas regarding professional autonomy. Internal tensions arise: while part of the team endorses algorithmic recommendations, others question them for insufficiently weighing contextual variables. Vanina and several colleagues reflect on the real scope of their autonomy vis-à-vis decisions suggested by the platform, even under supervision, and ask whether the mere presence of automated recommendations influences their creativity and the full exercise of professional judgement, particularly in urgent decision-making contexts.

The arrival of a placement student rekindles concerns about transparency. After observing the AI-driven workflow for several weeks, she asks her supervisor whether citizens should be explicitly informed that these tools are used in professional practice. Many service users are unaware that their personal data are processed by algorithms and do not grasp the implications of such use. The team considers providing an information-and-consent document but acknowledges that, even when signed, information asymmetries and uncertainty about future data uses would persist.

Incremental Potentials of AI

In this setting, AI yields modest but tangible gains: more efficient appointment management helps to reduce waiting times; automated transcription, minute-taking, and record-keeping relieve professionals of time-consuming administrative work,

freeing hours for direct, citizen-facing support; and draft generation for assessments and intervention plans provides structured starting points that practitioners can refine with contextual knowledge. These tools can support professional practice, improve continuity of care, and shorten response times; however, they require ongoing scrutiny, as many of their technical, ethical, and legal implications are not yet fully understood.

Emerging Epistemological Dilemmas

The analysis of emerging dilemmas reveals that the use of AI in social work is not limited to its impact in terms of efficiency or technical capacity but is deeply conditioned by the existence of transparent governance, qualified human oversight mechanisms, and a critical and vigilant approach to its application.

This scenario demands, more than ever, a robust critical education anchored in ethical and normative frameworks that deepen the commitment to social justice. Such a perspective begins with the guarantee of equitable access to services and social rights, while acknowledging territorial inequalities and the specificities of distinct socio-economic and cultural contexts.

It is therefore essential to critically examine the relationship between the efficiency gains promoted by AI and the potential risks these innovations may pose to professional autonomy, transparency in decision-making processes, and the realisation, or even regression, of social rights. In certain contexts, there is a risk that administrative bodies may use technological incorporation as a justification for reducing professional staff, promoting the indiscriminate use of AI without recognising the centrality of professional intervention grounded in relational work, qualified listening, and ethical commitment.

Strengthening this commitment requires integrating ethical reflection and digital literacy into professional training, ensuring that future practitioners are prepared to critically assess the implications of technological tools. It also depends on institutional policies that value ethical deliberation, promote collective accountability, and encourage the participation of social workers in the design and governance of AI systems. Equally important is the creation of collective spaces within social service organisations, such as professional supervision sessions, peer discussion groups, and continuing education initiatives, where practitioners can reflect on ethical dilemmas, share experiences, and develop critical perspectives on the integration of technology into their daily work. Beyond the institutional setting, professional associations and collective bodies also have a crucial role in fostering ongoing dialogue, advocacy, and ethical guidance for the responsible use of AI across the field.

AI Implications for Social Organisations and Decision-Making Processes

This section discusses how AI can improve transparency and efficiency in bureaucratic systems but may also deepen systemic inequities if not properly

regulated. Issues of accountability, algorithmic governance, and the centralisation of decision-making power are explored. It argues that organisations must develop AI literacy and ethical oversight frameworks to ensure that algorithmic decisions align with social justice values.

Hypothetical Scenario

AI-powered tools are increasingly used by social organisations to enhance service delivery and streamline case management. In the context of homelessness, machine learning systems promise more efficient resource allocation, predictive identification of high-risk cases, and improved inter-agency coordination. However, their integration into bureaucratic processes introduces new ethical risks. Algorithmic opacity, inadequate regulatory frameworks, and data-driven decision-making can inadvertently reproduce systemic inequities, disempower frontline workers, and erode the participatory ethos central to social work. This scenario examines the implications of AI in a housing-first organisation, raising urgent questions about justice, accountability, and human oversight.

In a large metropolitan area, a publicly funded social organisation—Redes Urbanas de Apoyo (RUA)—has begun using an AI-based decision-support system named Algoshelter to manage its services for people experiencing homelessness. Developed in partnership with a private tech firm, Algoshelter draws from a database combining internal records (e.g., case notes, service history), municipal data (e.g., police interactions, emergency shelter use), and health service usage to assess vulnerability scores and recommend intervention pathways.

The system classifies individuals into priority tiers, suggesting who should receive immediate housing placement, psychosocial support, or administrative follow-up. Social workers input updated notes via voice-to-text apps; algorithmic prompts guide assessment forms, automated dashboards flag urgent cases, and streamline weekly team reviews. Efficiency improves time spent on documentation drops by 50%, allowing more outreach visits.

Yet not all is well. The algorithm uses proxies for 'risk' that correlate with racialised policing patterns and penalise irregular service use: common among individuals with undocumented status or mental health conditions. Appeals to override recommendations require extensive justification and director approval. Social workers feel reduced to data-entry agents rather than ethical decision-makers. The algorithm's code remains proprietary, and the tech partner refuses external audits, citing intellectual property.

Experience on the Ground

Clara, a frontline outreach worker, initially welcomes the system. Algoshelter seems to relieve her from repetitive reporting, and the smart dashboard helps organise her daily visits. However, she notices troubling patterns: clients with complex, nonlinear trajectories are frequently deprioritised by the system. One woman, Iria, fleeing domestic violence without a formal ID or address history, scores low in

'housing readiness' despite evident need. The algorithm overlooks relational, situational, and trauma-informed criteria central to Clara's practice.

When Clara contests Iria's classification, the override request is delayed. By the time housing is approved, Iria has left the area. Clara feels complicit in a process that devalues professional judgement. During a team meeting, she raises concerns, but supervisors urge compliance, citing funding requirements linked to efficiency metrics.

Meanwhile, clients begin to suspect that 'the computer decides everything'. Trust erodes. Clara starts keeping parallel notes outside the system, an act of resistance and care, but one that violates policy.

Incremental Potentials of AI

Despite the dilemmas, Algoshelter shows measurable gains. It identifies patterns across large datasets that would escape manual review, such as early indicators of eviction risk among shelter residents. Administrative tasks are reduced, and managers report faster triage decisions. Predictive alerts have, on occasion, enabled timely intervention and hospitalisation avoidance.

These benefits, however, remain incremental and uneven. Like human systems, AI tools are shaped by the conditions of their use, ethical judgement, professional discretion, and even informal workarounds. The AI system improves bureaucratic throughput, but not necessarily structural outcomes. Its most effective use occurs when embedded in teams with strong ethical cultures, transparent processes, and space for critical dialogue.

Emerging Governance Dilemmas

This scenario reveals deep dilemmas at the intersection of AI, care, and social justice. Algorithmic governance, framed as neutral and efficient, conceals normative assumptions and opaque exclusions. When decisions are centralised in proprietary code, the professional autonomy of social workers erodes. The *deskilling* of practice and the *delegitimisation* of situated knowledge risk displacing the relational and ethical dimensions of social work.

Furthermore, accountability mechanisms remain weak: Who answers when a system misclassifies? How is harm redressed? Transparency is limited, with no external validation of models or data sources. Clients lack meaningful appeal channels, and their data (often gathered under distress) feeds systems that shape their futures without consent or comprehension.

Such risks are exacerbated by regulatory lag, where public institutions lack the legal, ethical, and technical capacity to oversee AI. Without clear frameworks, social organisations face mounting pressure to deliver metrics, not justice.

Conclusions

This chapter has examined the intersection of AI and social work, exploring both the transformative potential and the ethical, political, and practical risks associated

with AI technologies in this field. While AI offers innovative tools that may augment decision-making, improve service delivery, and generate new knowledge, it also poses significant challenges to core social work values, such as social justice, human rights, and relational practice. These objectives have been addressed through the analysis of four different scenarios, each illustrating how AI can simultaneously enable opportunities and generate tensions that social work must critically engage with.

The four scenarios underscore an urgent need for AI literacy among social workers and ethical oversight frameworks within organisations. Training must enable professionals to critically interpret algorithmic outputs, understand systemic biases, and advocate for clients whose experiences elude datafication. Ethical review boards, interdisciplinary audits, and participatory design processes should be institutionalised. Strengthening this ethical infrastructure also requires that social workers participate as co-designers in the development and adaptation of digital tools, ensuring that these systems reflect professional knowledge, contextual understanding, and human values rather than purely technical efficiency.

In line with social work's commitment to human dignity and structural change, AI must be interrogated not only for what it does, but for how and for whom it does it**.** A critical warning is therefore necessary: AI systems cannot replace the reflective, value-based, and context-sensitive judgement that defines the profession. To mitigate these risks, it is essential to strengthen transparent governance, ensure continuous professional education, and equip practitioners with competencies in digital literacy, ethical reasoning, and critical reflection. Equally important is the preservation of professional supervision, peer dialogue, and collective reflection spaces within organisations, arenas where practitioners can question, analyse, and deliberate about technological practices and their ethical implications.

The four scenarios analysed reveal that the integration of AI into social work unfolds as an ambivalent horizon**,** marked by opportunities and risks, promises and dilemmas. This horizon, between the lighthouse and the storm, symbolises both the illuminating potential of technology and the disruptive forces that threaten to unsettle the ethical and relational foundations of the profession. Regarding education and training, AI broadens access and enables personalised learning pathways, yet it also threatens students' intellectual autonomy, reinforces cultural biases, and may deepen the digital divide. The light lies in pedagogical innovation; the shadow, in the commodification of knowledge and surveillance. Regarding research, algorithms facilitate the analysis of vast datasets and the detection of patterns, strengthening applied research. At the same time, they risk displacing qualitative and narrative methodologies, reproducing hegemonic epistemologies. The light emerges in expanded analytic capacity; the shadow, in the loss of plural and in-depth interpretation. Regarding professional intervention, automation relieves administrative burden and accelerates procedures, improving responsiveness. Yet it may erode the human relationship at the heart of practice, undermine professional autonomy, and justify staff reductions. The light appears in efficiency and continuity of care; the shadow, in standardised diagnoses and the fragility of social rights. Finally, regarding organisations and decision-making processes, AI

promises greater transparency, traceability, and speed in management, but risks concentrating decision-making in opaque systems and amplifying structural inequalities. The light lies in inter-agency coordination and resource optimisation, the shadow, in opacity, loss of trust, and the weakening of ethical deliberation.

Taken together, these scenarios show that the incorporation of AI in social work is not an inevitable destiny but rather a contested field. Its consequences will depend on the political, institutional, and ethical frameworks guiding its deployment. The challenge is not to reject or embrace technology uncritically but to strengthen digital literacy, secure robust democratic regulation, and reaffirm social work's foundational commitment to social justice, human dignity, and collective rights. In sum, AI presents both a challenge and an opportunity for social work. Rather than offering prescriptive answers, it advocates for a critical, reflective stance that embraces uncertainty while striving for social equity and human dignity. It emphasises the need to develop adaptive frameworks, ethical guidelines, and participatory models of innovation that align technological development with the core mission of social work: human rights, inclusion, dignity, and social justice. Ultimately, the future of social work in relation to AI unfolds, as the title of this chapter suggests, *between the lighthouse and the storm***:** between the light of innovation that guides new horizons and the ethical turbulence that tests the humanistic core of the profession. Navigating this ambivalent horizon will require keeping a steady course guided by the principles of justice, empathy, and responsibility that define the essence of social work.

Note

1 Research funded by Ministerio de Ciencia, Innovación y Universidades Code: PID2023–150587OB-I00 ADOPCION DE LA TECNOLOGIA, EXPERTOS INFORMALES Y EDUCACION DIGITAL EN ADULTOS MAYORES.

References

Arora, A., Barrett, M., Lee, E., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization, 33*(3), 100478. https://doi.org/10.1016/j.infoandorg.2023.100478

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, *32*, 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Brandao, P. R. (2025). The impact of artificial intelligence on modern society. *AI*, *6*(8), 190. https://doi.org/10.3390/ai6080190

Bustio-Martínez, L., Herrera-Semenets, V., Feregrino-Uribe, C., & Van der Berg, J. (2024). Making adequate use of generative language models in scientific research. ***Available at SSRN*:** https://ssrn.com/abstract=5367076 or https://doi.org/10.2139/ssrn.5367076

Černý, M. (2024). *AI literacy: Concepts, approaches and open questions*. Masaryk University. https://www.muni.cz/en/people/268947-michal-cerny/publications

Dritsas, E., & Trigka, M. (2025). Machine learning and data science in social sciences: Methods, applications, and future directions. *IEEE Access*, *13*, 105334–105352. https://doi.org/10.1109/ACCESS.2025.3578906

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144, and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union, L 168/1*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Garkisch, M., & Goldkind, L. (2025). Considering a unified model of artificial intelligence enhanced social work: A systematic review. *Journal of Human Rights and Social Work, 10*, 23–42. https://doi.org/10.1007/s41134-024-00326-y

Junqueira, E. (2025). Inteligência artificial generativa e estudantes universitários no contexto dos multiletramentos. *Revista Docência e Cibercultura, 9*(1), 1–17. https://doi.org/10.12957/redoc.2025.81548

Nayir, F., Sari, T., & Bozkurt, A. (2024). Reimagining education: Bridging artificial intelligence, transhumanism, and critical pedagogy. *Journal of Educational Technology & Online Learning*, *7*(1), 102–115. https://doi.org/10.31681/jetol.1308022

Papaspyridis, A., & La Greca, J. (2023). AI and education: Will the promise be fulfilled? In D. Araya & P. Marber (Eds.), *Augmented education in the global age: artificial intelligence and the future of learning and work* (pp. 119–136). Routledge.

Peponakis, M., Kapidakis, S., & Doerr, M. (2024). From calculations to reasoning: History, trends, and the potential of computational ethnography and computational social anthropology. *Social Science Computer Review*, *42*(1), 84–102. https://journals.sagepub.com/doi/pdf/10.1177/08944393231167692

Reamer, F. (2023). Artificial intelligence in social work: Emerging ethical issues. *International Journal of Social Work Values and Ethics*, *20*(2), 52–71. https://doi.org/10.55521/10-020-205

Riley, S., Healy-Cullen, S., Rice, C., Tiidenberg, K., Hawkey, A., Evans, A., Stephen, C., Tappin, J., Ensslin, A., & Morison, T. (2025). Problematising menstrual tracking apps: Presenting a novel critical scoping review methodology for mapping and interpreting research literature. *Psychology & Health,* 1–25. https://doi.org/10.1080/08870446.2024.2445518

Toli, L., & Manasa, G. M. (2024, abril). Artificial intelligence: Opportunities and challenges for the social education and profession. *Scholars Bulletin, 10*(4), 143–147.

Vesna, L., Sawale, P. S., Kaul, P., Pal, S., & Murthy, B. S. N. V. R. (2025). Digital divide in AI-powered education: Challenges and solutions for equitable learning. *Journal of Information Systems Engineering and Management, 10*(21s), Article 3327. https://doi.org/10.52783/jisem.v10i21s.3327

Wang, H., Tlili, A., Huang, R., Cai, Z., Li, M., Cheng, Z., Yang, D., Li, M., Zhu, X., & Fei, C. (2023). Examining the applications of intelligent tutoring systems in real educational contexts: A systematic literature review from the social experiment perspective. *Education and Information Technologies, 28*, 9113–9148. https://doi.org/10.1007/s10639-022-11555-x

18 The Use of AI-Based Counselling

Professional and Ethical Implications for Social Work

Monika Alamdar-Niemann and Sabrina Heinl-Vako

AI-Based Counselling in Social Work: Potentials, Risks, and Professional Considerations

At present, artificial intelligence (AI)–based counselling is already being applied in different fields. Not only are AI-based chatbots perceived very positively by users (Ayers et al., 2023, p. 3), but initial studies also indicate their high effectiveness (Durden et al., 2023, p. 1). Given these findings, as well as the severe shortage of qualified professionals in counselling and the simultaneously high demand for psychosocial support, it can be assumed that the use of AI-based chatbots will increase significantly soon. In this context, it is important to note that the programmes developed to date are mostly based on the mechanisms of cognitive behavioural therapies (CBTs) and are therefore more psychological in nature (Fulmer et al., 2018, p. 1). If these were to be applied in the field of social work, the question arises as to whether they would then have to be programmed in accordance with the core values of the social work profession. In this context, it is important for social work scientists to be involved in the programming and research of the programmes.

On closer inspection, it becomes apparent that the use of AI-supported counselling in the field of social work could have both positive and negative potential effects. Above all, there is a risk that programmes will not be developed in collaboration with social workers themselves, meaning that potential negative effects on users will not be sufficiently considered.

These implications range from questions concerning the bots' capacity for 'professional' action to issues of data protection. If chatbots are used within social work, it is crucial that they do not overshadow or undermine the ethical and theoretical foundations of the profession. Instead, they should support social workers in pursuing and strengthening these core values more effectively in their practice. In this chapter, we therefore examine some of the potential impacts of using AI-supported chatbots on clients, professionals, educators, and students in social work, and outline recommendations for practice within this context.

DOI: 10.4324/9781003627920-18
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

Current Expert Discourse on AI-Supported Consulting

The use of AI is a central topic of research in current professional discourse and is already being applied in practice in isolated cases. For example, there are private providers such as 'Therapeak' in Germany (Therapeak, 2025), which already offer fully AI-supported counselling options, or Woebot from the US (Woebot Health, 2025b), which can also be used by people who download and use the associated app. These programmes communicate with users in writing. Woebot was founded by clinical research psychologist Dr. Alison Darcy in 2017 (Woebot Health, 2025b) and gives 24/7 support as an early intervention tool for clients who need support for moderate issues. Sarp and Nisan Derinsu, a psychologist and a software engineer founded Therapeak (Therapeak, 2025) with the same intention of offering psychosocial support for users, in the form of fully-AI-based counselling programmes.

Tess, a psychological AI chatbot developed by X2AI Inc. (now Cass, 2025), interacts with users via text-based platforms such as SMS and Facebook Messenger to deliver psychoeducational materials, coping mechanisms, and emotional support (Fulmer et al., 2018, pp. 3–4). Although not intended to replace conventional psychotherapy, Tess serves as a supplementary or self-guided tool. The app is also available on the open market and has been tested for effectiveness in studies. It was shown that it was able to reduce symptoms of depression and anxiety by communicating with its users and applying CBT-based interventions (Fulmer et al., 2018, p. 11).

There are also programmes such as Replika[1] or Cleverbot[2] which, although not based on psychological or psychotherapeutic concepts such as CBT, still offer users a general opportunity to communicate via chat. Several studies conducted by scientists at Harvard University show that using these programmes can lead to a reduction in feelings of loneliness (De Freitas et al., 2024, p. 1). These programmes were not developed by people from the psychosocial field, such as psychologists, but are freely available programmes that can communicate with people and are used for this purpose.

It appears that fully AI-supported counselling programmes are not yet being used in the field of social work. However, they are being used in social work education so that students can learn counselling skills (Lippert et al., 2024, p. 41). In these cases, the applications were programmed by the people involved in the research project. For example, there is the 'ViKL' (Virtual Client) project, in which an AI-supported chatbot is used in teaching students so that they can develop their counselling skills (Lippert et al., 2024, p. 41). In addition, there are numerous studies, mainly in the field of psychology, which investigate the effects of AI-supported counselling on people and the effectiveness of using these programmes. In various studies (see Eichenberg & Küsel, 2016, p. 94), it has been shown that different forms of online counselling are effective and, in certain situations—such as with topics that involve shame—even significantly more accessible than in-person services (see Zizelmann, 2023, p. 99). Until now, research and applications have mainly focused on text-based interactions with AI. Further developments in robotics would be necessary to enable AI-supported counselling in person (Lehmann, 2025, S. 117).

Regarding the integration of AI in social work, Wang (2025) explores the convergence of AI and social work, outlining both the transformative potential and the complex challenges associated with this integration. Wang advocates for the development of comprehensive ethical and regulatory frameworks; increased technological competency among social work professionals, the so-called 'technical literacy' (Wang, 2025, p. 267); equitable access to AI infrastructure; and the design of hybrid human–AI interaction models.

As AI-based social work counselling is the main topic here, in general the innovative character of AI for counselling tasks in social work must be focused under the empathetic intelligence. It refers to enhancing AI's capacity to detect and interpret human emotions, react in a suitable manner, and affect others' emotional states. To do this effectively, AI must develop skills in communication, social interaction, and relationship management. These abilities are essential in roles that involve close personal engagement and emotional sensitivity, such as those of therapists or social workers (Huang & Rust, 2018, p. 159). On top of that, it is important, that an AI-based counselling chatbot can behave 'professionally' and act according to the central core ethical values of the profession of social work. These standards are not established by software developers; rather, they are defined by official social work organisations representing experts and practitioners who delineate the core values of the profession. The ethical guidelines of these developers, which encompass subjects such as respect, human rights, fairness, data protection, and responsibility, are made publicly available on the websites and in the documents of the professional associations (IASSW, 2025; IFSW, 2018). Li et al. (2023) analysed 35 experimental investigations, including 15 randomised controlled trials (RCTs), that examined the use of AI-powered conversational agents (CAs) designed to support mental health and well-being. Regarding the AI-based counselling, there is evidence that Chatbots or CAs show moderate to large effects in reducing symptoms of depression and stress neurosis (Li et al., 2023, p. 9).

Beatty et al. (2022) provide empirical support for the hypothesis that therapeutic relationships can be formed between individuals and CAs in mental health contexts. Their study examined Wysa, an AI-driven mental health application designed to create a supportive virtual space for user-initiated interactions. The platform enables AI-mediated active listening and emotional support, while also offering self-help tools and evidence-based interventions, including those based on CBT. Wysa's self-help tools are designed by its in-house team, which incorporates designers and mental health professionals. However, these tools only become effective when users engage with them regularly. The authors note limitations in the dataset and highlight the need for control groups to enhance comparative analysis (Beatty et al., 2022, p. 6).

Existing literature frequently underrepresents or insufficiently documents user involvement in the design of mental health chatbots, particularly in terms of user-centred adaptation and participatory development processes (Algumaei et al., 2025, pp. 22–26). It also becomes apparent that there is little social work research, as opposed to psychological research, on the development and evaluation of chatbots, which is why it is not yet entirely clear how chatbots can and should

be designed for social work. This review underscores the considerable potential of AI-driven mental health chatbots to provide scalable and personalised support. Nevertheless, significant limitations persist, most notably in the areas of ethical considerations, cultural responsiveness, inclusivity, and empirical substantiation. Technological advancements, while promising, do not inherently ensure clinical efficacy or foster public trust. A notable gap in the current research landscape as well is the underrepresentation of the Global South, which restricts the applicability and cultural relevance of findings. To ensure equitable and inclusive development, future studies must actively incorporate perspectives and data from low- and middle-income countries, social workers and their clients.

Potential Implications of the Use of AI-Supported Counselling in Social Work

The use of AI-supported counselling chatbots could have implications for both clients and professionals in social work. Therefore, social workers and clients must continuously engage in critical reflection on the possible consequences of such use, assessing them in a way that ensures a safe and responsible application aligned with the core values of the social work profession. In this context, questions regarding the safety and quality of AI-supported chatbot applications naturally arise.

The Potential Impact of AI-Supported Counselling on Its Users within the Context of Social Work

At present, there are already AI-supported counselling chatbots available. To use these, users must pay for the service themselves. There are no reports yet of the widespread integration of bots in social work. The programmes that are currently in use are generally only capable of offering counselling in written form, i.e., in the form of chats (Lehmann, 2025, p. 118).

The potential negative and positive effects on users depend heavily on the design and quality of the bot itself, as well as on the setting in which it is used. As already mentioned, there are programmes that are conceptualised by experts and based on psychological approaches. In contrast, there are freely available programmes that are only capable of general communication with users.

Potential negative implications of using large language models to communicate in psychosocial crises can be illustrated using programmes such as ChatGPT. It is questionable whether ChatGPT is suitable for psychosocial counselling due to the risks described below if the programme is used for counselling purposes. Nevertheless, it is sometimes used in crises when users see no alternative option because places in counselling centres and psychotherapy practices are limited (Woebot Health, 2025b) or because they fear discrimination and stigmatisation if they seek psychotherapy (Carlbring et al., 2023, p. 2). AI-supported counselling chatbots that could be used in health and social services must be designed in such a way that the likelihood of the risks described below can be minimised.

If models such as ChatGPT are used nonetheless, there is a risk for users that signs of suicidal tendencies or indications of child welfare risks will not be recognised and that these situations will not be handled professionally. This became clear in the tragic case of a Californian teenager who turned to ChatGPT in a suicidal crisis and received advice about his suicide plans. The crisis could not be averted and the teenager committed suicide (Yousif, 2025). AI-supported counselling chatbots would have to be programmed in such a way that they can act appropriately in such cases. For example, in the event of a suicidal crisis, psychiatric help should be requested in the form of an emergency medical response. If there are indications of child welfare risks during a consultation, legally prescribed action is also required, which in acute cases may include reporting the matter to child protective services. AI-supported counselling chatbots would have to be capable of initiating such further interventions to eliminate or minimise the acute and potential danger to the lives of the persons involved. The question currently remains as to how this can be implemented, as to date no programmes have been described in the literature that would be capable, for example, of calling an ambulance or initiating a report to the youth welfare office in the cases described.

However, developers of Woebot point out that sensitive conversation content is monitored so that they can respond in such situations (Woebot Health, 2025a). The question arises as to how, in emergencies arising from the counselling situation, it can be ensured that the AI initiates further assistance itself, for example, by calling an ambulance or notifying the youth welfare office. Checking conversation logs, on the other hand, could potentially cause delays and prevent adequate, timely intervention. This raises the question of whether checking conversation logs creates a similarly high workload, which could make it more efficient and safer for professionals to conduct the counselling themselves.

When ChatGPT is used for communication, there is also the problem that personal data are not protected (Ruschemeier, 2025, p. 349). This is particularly critical when it comes to disclosing personal matters, especially in the case of crisis. If AI-supported counselling chatbots were used in the field of social work, they would have to be programmed in such a way that users' personal data can be processed in accordance with nationally applicable data protection guidelines. It must be conclusively clarified what happens to the data collected in this context. Clients need to know when their own data is collected and where and for how long it must be stored (Brandtzæg et al., 2021, p. 11).

There is also a risk that the chatbots could generate stereotypes and reinforce social power structures. Schmidt and Thiessen refer to the stereotypes generated by chatbots when they state that the use of ChatGPT for communication in crisis can lead to gender stereotypes and sexist assumptions being reproduced without reflection. For example, AI often implicitly refers to a female addressee, which can be seen in the one-sided recommendation of women-specific offers of help. Male options only become visible through targeted enquiry. In addition, the answers reflect traditional role patterns: female figures appear emotional and vulnerable, men as dominant or aggressive. Ambivalences, social differentiation, or norm-critical perspectives are absent, which perpetuates existing power relations (Schmidt & Thiessen, 2024, p. 80). Counselling chatbots used in the context of

social work would have to be programmed in such a way that stereotypes are broken down rather than generated, so as not to reinforce social power relations and to enable a diversity-sensitive way of working.

Currently, there is no legal requirement that AI counselling chatbots may only be programmed with the involvement of professional specialists. This could lead to models not being developed in line with the principles of various professions, such as social work, psychology, or psychotherapy, and could increase the risks described above for users. The EU AI Act stipulates that AI must not be used when working with vulnerable groups if there is a risk of harm (Dötterl, 2025, p. 223) and that the services offered should be designed based on medical principles (Dötterl, 2025, p. 226). Yet the explanations in the EU AI Act remain relatively vague. This raises the question of how the market in this area can be adequately regulated.

What is certain, however, is that marketing AI-supported advice could generate high returns, regardless of how the counselling chatbots are programmed. This becomes clear when it is considered that AI-supported counselling programmes can be offered around the clock, at any time of the day or night. The bots don't need vacations, never get sick and can be used by numerous users at the same time. It becomes clear what high profit margins could be possible when offering AI-supported counselling programmes commercially. In this context, further regulations will be needed to reduce the risk of the dangers to users described above.

On the other hand, AI-supported counselling chatbots are perceived by users as 'empathetic' and helpful. This is indicated, for example, by a study by Ayers et al., in which the responses of doctors were compared with those of a chatbot. Users perceived the doctors' responses as empathetic or very empathetic 4.6% of the time, while the chatbot was rated as such 45.1% of the time. The chatbot's responses were also rated higher in terms of quality (Ayers et al., 2023, p. 3). Skjuve et al. conducted interviews with people who communicated with the chatbot 'Replika' over and concluded that users rated their interaction with the chatbot positively and that, from their perspective, their well-being was improved by communicating with the bot (Skjuve et al., 2021, p. 1).

The findings of further empirical studies show that AI-supported interactions are not only perceived as valuable and helpful, but that they are also highly effective. For example, evaluations of Woebot show that using the programme over a period of eight weeks can contribute to a targeted reduction in symptoms of burnout and stress (Durden et al., 2023, p. 1).

In this context, it becomes clear that the use of AI-supported counselling chatbots can be highly effective and can counteract the shortage of skilled workers and the supply crisis in the field of psychosocial counselling. However, to avoid negative implications, high standards should be pursued and monitored in the design of the programmes.

Potential Impact of the Implementation of AI-Supported Counselling on Social Work Professionals

The use of AI-supported counselling also has potential implications for skilled workers. Although the acute shortage of skilled workers is likely to prevent social

workers from being completely replaced by AI-supported chatbots, there is at least a risk that the number of jobs will be reduced to a minimum and that employers will prefer AI-supported counselling bots to employees, because the use of a chatbot for companies is not accompanied by the obligation to regularly pay an appropriate salary or to ensure high-quality working conditions. This may have a negative impact on the already precarious conditions in social work.

In addition, the focus on the core values of the respective professions may be lost, as chatbots may not embody these values in practice and may not be able to incorporate them into their actions. Some of the fundamental values of social work, such as showing solidarity, have a clear socio-political aspect, which is promoted by professionals not only as representatives of a profession, but also as politicised individuals in society. The Professional Association of Social Work in Germany points out that professionals should demonstrate solidarity to stand up for the goals and rights of different groups in solidarity with others (German Professional Association for Social Work, 2024, p. 33).

Counselling chatbots do not represent politicised individuals in society. The question arises as to whether the promotion of solidarity with clients could be programmed into the 'professional identity' of a bot as a professional ethical value and how this would be reflected in the bot's actions. In addition, social workers should work to drive socio-political change that promotes a humane society through political and methodically targeted action (German Professional Association for Social Work, 2024, p. 35). In this context, there is hope that the use of AI-supported counselling chatbots could lead to social workers having more time for the policy work described above (Dumman, 2025, p. 53). However, there is currently no empirical evidence that such time savings can be achieved using AI-supported counselling (Dummann, 2025, p. 51)

Furthermore, social work comprehends itself as a human rights profession. As one of the most important advocates of social work as a human rights profession in German-speaking countries, Staub-Bernasconi refers to the International Association of Schools of Social Work when she emphasises that social workers should be committed to upholding the human rights of their clients (Staub-Bernasconi, 2007, p. 9 f.). Dumman also points out that a human rights-based orientation must not be supplanted using AI (Dumman, 2025, p. 54).

AI-supported counselling chatbots that could be used in social work would have to be programmed in such a way that they have a professional identity and the ability to act professionally, according to the ethical and theoretical core of social work. For this reason, the literature points out that social workers should be involved in the programming of AI-supported counselling chatbots (Macsenaere & Feist-Ortmanns, 2025, p. 95).

Professional and Ethical Implications of Using AI-Supported Counselling for Social Work and Recommendations for Action

The potential effects described point to various professional and ethical implications and recommendations for action for social work in the context of using AI-supported counselling in the future. In general, professionals, students, and

teachers in social work should take a critical approach to the potential use of AI in general and in the context of counselling in particular, as this area is set to expand significantly in the coming years (Lehmann, 2025, p. 118).

Social work professionals and students must be able to critically assess the quality and applicability of AI-supported counselling programmes, once they will be used in practice. They should specifically ask whether the use of a specific programme is necessary and appropriate in the individual context, or whether it is more likely to cause harm than benefit. This is also in line with the EU AI Act, which states that AI must not be used with vulnerable groups if it poses a potential serious risk (Dötterl, 2025, p. 223). They should not lightly advocate for its use, even if it is already being used in various areas, but rather critically reflect on the potential risks of its application.

To do this, professionals need to know what questions to ask about AI-supported counselling programmes. They should question how the programme responds in crises. Is there a possibility that the programme will contact professionals directly or even call emergency services in the event of suicidal statements or serious indications of child welfare risks? Do conversations have to be reviewed retrospectively and, if so how soon after the counselling session does this review take place?

Students and professionals must also question the data protection compliance of the programmes and report back if the application cannot meet the requirements of the General Data Protection Regulation (GDPR) to protect the data of the addressees. In addition, they must question whether the programmes used generate stereotypes and how this could potentially affect the addressees. They should make it clear that the programme must be designed in such a way that the generation of stereotypes can be ruled out or at least greatly reduced.

It is therefore of utmost importance that social work researchers, social workers, and students find ways to participate in or initiate research projects on the development and investigation of AI-supported counselling chatbots. At present, such projects tend to be carried out by psychologists who develop and research psychosocial counselling programmes, usually in collaboration with software developers. Particularly considering the potential cost savings associated with the use of bots, there is a risk that such programmes will be developed for social services without involving social workers. This raises the danger that the programmes will not be developed in line with the ethical and professional principles of social work. Therefore, social workers should always question whose interests are being focused on in this context and advocate that the well-being of the recipients must always be given greater weight than efficiency and profit increases for the providers. The use of AI in health and social services must also be considered from a labour law and trade union policy perspective. Students and professionals in social work should advocate for AI to be used to improve the situation of employees.

The use of AI should never be allowed to worsen the conditions under which social work is carried out. Therefore, professionals in the field of social work should constantly reflect critically on the use of AI and its possible consequences in practice. To do this, they need to be familiar with the current state of research on the topic and discuss it in team meetings. The results of this critical reflection

should be considered when planning the implementation of AI-supported chatbots in the institution. In this context, networking with professional associations of social work on a national and international level and trade unions is also essential in order to be able to take action against such developments at an early stage.

Once AI-based counselling chatbots are in use in the field of social work, professionals should use the time resources freed up by the use of AI to engage in policy work and fulfil the political mandate of social work. Potential socio-political changes resulting from the use of AI in counselling should always be critically reflected upon within the team.

In addition, it is important to consider how the use of AI counselling chatbots could affect professional conduct within the team. The bot should be programmed in such a way that the professional ethical values of social work are always considered. In essence, it would even be necessary for the bot to have the professional 'identity' of a social worker and the ability to act professionally.

In this context, social work researchers and teachers are encouraged to promote interdisciplinary projects in which AI-supported counselling chatbots are programmed jointly so that, on the one hand, they embody the professional identity and ability to act professionally, in the sense of a social worker, in the programme. On the other hand, the effects of using AI-supported counselling in practice must always be evaluated and researched to leverage positive effects and limit negative consequences. This is where social work research plays a crucial role.

Teachers should therefore encourage students to conduct research in this area and, for example, write theses on the subject. It is also very important to promote young academics in relevant social work research projects. In this context, specific professorships for digitalisation and AI in the context of social work should be established at universities. Social work curricula should include relevant modules.

Conclusion

The use of AI-supported chatbots is expected to increase due to the severe shortage of qualified professionals in the field, coupled with a high demand for psychosocial counselling. AI-based chatbots already exist and are applied across different areas and settings. As of now they are mostly used commercially and used by individual users, who pay for the service. They are often based on concepts such as CBT or other psychological theories, but it is possible, that they will be used more widely in the field of social work as well. Current findings indicate that users perceive these bots as 'empathetic' and respectful. Furthermore, the high effectiveness of many of these programmes is now well documented. Nevertheless, challenges can arise, especially when these programmes are not created with professionals out of the field of social work itself. For example, in the areas of data protection and the 'professional' handling of crises. There is also a risk that these programmes reinforce stereotypes. It is likely that these programmes will be developed without the involvement of social workers or other professionals in the health and social sector. Therefore, it is essential that professionals, educators and students in this

field actively engage to ensure that future programmes operate in accordance with the ethical standards and theoretical frameworks of social work. Within social work teams, the use of AI-supported chatbots should always be critically reflected upon. Social workers should strive to ensure that the use of such programmes strengthens and promotes core professional values and ethical practice rather than undermining them.

Future investigations should prioritise the co-creation of culturally and contextually sensitive systems, with a focus on embedding chatbot-based interventions seamlessly into everyday life and existing healthcare infrastructures. Longitudinal research and rigorous evaluation frameworks are essential to bridging the current divide between technological innovation and empirical evidence. Moreover, AI systems must be developed and governed with transparency and ethical oversight, emphasising privacy protection, interpretability, and equitable access—particularly for populations that face digital exclusion. This includes the design of multilingual and accessibility-conscious interfaces.

As emotion-recognition capabilities in AI continue to evolve, the establishment of robust safety mechanisms and escalation protocols is critical. While chatbots may serve as valuable adjuncts to mental health care, especially in resource-limited settings, they are not substitutes for human clinicians. Maintaining clear distinctions between automated support and professional therapeutic care is therefore essential.

Notes

1 Cleverbot is a chatbot, which communicates with its users via chat. Users can download an app to use the application. https://www.cleverbot.com/.
2 Replika is also a chatbot and a program, in which users can create an avatar and a "replica" of themselves. The website describes it as an "AI companion who cares": www.replika.com.

References

Algumaei, A., Yaacob, N. M., Doheir, M., Al-Andoli, M. N., & Algumaie, M. (2025). Symmetric therapeutic frameworks and ethical dimensions in AI-based mental health chatbots (2020–2025): A systematic review of design patterns, cultural balance, and structural symmetry. *Symmetry, 17*(7), 1082. https://doi.org/10.3390/sym17071082

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Christopher, A., Longhurst, M., Hogarth, M., & Smith, D. M. (2023, April). *Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine.* https://doi.org/10.1001/jamainternmed.2023.1838

Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022, April 11). Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study. *Frontiers in Digital Health, 4,* 847991. https://doi.org/10.3389/fdgth.2022.847991

Brandtzæg, P. B., Skjuve, M. B., Dysthe, K. K., & Følstad, A. (2021, April). When the social becomes non-human: Young people's perception of social support in chatbots. In *Conference paper presented at CHI 2021.*

Calbring, P., Hadjistavropoulos, H., Kleiboer, A., & Andersson, G. (2023). A new era in Internet interventions: The advent of Chat-GPT and AI-assisted therapist guidance. *Internet Interventions, 32*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10235420/pdf/main.pdf

Cass. (2025). Cass empowers healthcare providers and those who suffer from stress to achieve better health and wellness and save money too. https://www.cass.ai/anxiety

De Freitas, J., Uguralp, A. K., Uguralp, Z. O., & Puntoni, S. (2024). *AI companions reduce loneliness*. Working Paper 24-078. Harvard Business School. https://www.hbs.edu/ris/Publication%20Files/24-078_a3d2e2c7-eca1-4767-8543-122e818bf2e5.pdf

Dötterl, S. (2025). EU AI Act und Soziale Arbeit: Die KI-Verordnung und ihre Auswirkungen. In G. Linnemann, J. Löhe & B. Rottkemper (Hrsg.), *Künstliche Intelligenz in der Sozialen Arbeit: Grundlagen für Theorie und Praxis* (S. 221–140). Beltz Juventa Verlag.

Dumman, J. (2025). Bedeutung von KI für Disziplin und Profession der Sozialen Arbeit. In G. Linnemann, J. Löhe & B. Rottkemper (Hrsg.), *Künstliche Intelligenz in der Sozialen Arbeit: Grundlagen für Theorie und Praxis* (S. 47–59). Beltz Juventa Verlag.

Durden, E., Maddison, C., Pirner, S. J., Rapoport, A., Williams, A., Robinson, V. L., & Forman-Hoffman, V. L. (2023). *Changes in stress, burnout, and resilience associated with an 8-week intervention with relational agent "Woebot". Internet Interventions, 33* https://doi.org/10.1016/j.invent.2023.100637

Eichenberg, C., & Küsel, C. (2016). Zur Wirksamkeit von Online-Beratung und Online-Psychotherapie. *Resonanzen – E-Journal für biopsychosoziale Dialoge in Psychosomatischer Medizin, Psychotherapie, Supervision und Beratung, 4*(2), 93–107.

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), e64.

German Professional Association for Social Work. (2024). *Berufsethik des DBSH. Ethik und Werte in der Sozialen Arbeit.* https://www.dbsh.de/media/public/dbsh-bund/Profession/2025-05-21_DBSH_Berufsethik_2._Fassung.pdf

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research, 21*(2), 155–172. https://doi.org/10.1177/1094670517752459

IASSW (2025). Global social work statement of ethical principles IASSW https://www.iassw-aiets.org/wp-content/uploads/2018/04/Global-Social-Work-Statement-of-Ethical-Principles-IASSW-27-April-2018-1.pdf

IFSW (2018). Global social work statement of ethical principles. https://www.ifsw.org/global-social-work-statement-of-ethical-principles/

Lehmann, R. (2025). KI in der Beratung. In G. Linnemann, J. Löhe & B. Rottkemper (Hrsg.), *Künstliche Intelligenz in der Sozialen Arbeit: Grundlagen für Theorie und Praxis* (S. 117–127). Beltz Juventa Verlag.

Li, H., Zhang, R., Lee, Y.C., Kraut, R.E., & Mohr, D.C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine, 6*, 236. https://doi.org/10.1038/s41746-023-00979-5

Lippert, C., Rudolph, E., Poltermann, A., Engert, N., Lehmann, R., & Albrecht, J. (2024). Generative KI in der beraterischen Ausbildung. Der Einsatz eines*r virtuellen Klient*in als Übungstool für angehende Onlineberater*innen. *Zeitschrift für Onlineberatung und computervermittelte Kommunikation, 1*(3), 41–60.

Macsenaere, M., & Feist-Ortmanns, M. *KI in der Kinder- und Jugendhilfe* (S. 90–101). *Künstliche Intelligenz in der Sozialen Arbeit: Grundlagen für Theorie und Praxis* (S. 156–173). Beltz Juventa Verlag.

Ruschemeier, H. (2025). *LLMs und Datenschutzrecht. DuD – Datenschutz und Datensicherheit, 6,* Schwerpunkt.

Schmidt, J., & Thiessen, B. (2024). In_visible gender? Geschlechterkritische Reflexionen beim Beratungslernen mit digitalen Medien. *Journal Netzwerk Frauen- und Geschlechterforschung NRW,* (55), 77–83. https://duepublico2.uni-due.de/servlets/MCRFileNodeServlet/duepublico_derivate_00082283/Journal_Netzwerk_FGF_55_77.pdf

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzæg, P. B. (2021). My chatbot companion – A study of human–chatbot relationships. *International Journal of Human-Computer Studies, 149,* 102601. https://doi.org/10.1016/j.ijhcs.2021.102601

Staub-Bernasconi, S. (2007). *Soziale Arbeit: Dienstleistung oder Menschenrechtsprofession? Zum Selbstverständnis Sozialer Arbeit in Deutschland mit einem Seitenblick auf die internationale Diskussionslandschaft.*

Therapeak (2025). Über uns. https://therapeak.com/de/about

Wang, W. (2025). The integration of artificial intelligence and social work: Opportunities, risks, and future directions. *Frontiers in Humanities and Social Sciences.* https://doi.org/10.54691/hx3e5z06

Woebot Health. (2025a). *Technically advanced. Responsible AI.* https://woebothealth.com/technology-overview/

Woebot Health. (2025b). *Mental health needs have multiplied. Support hasn't.* https://woebothealth.com/

Yousif, N. (2025). *Parents of teenager who took his own life sue OpenAI.* https://www.bbc.com/news/articles/cgerwp7rdlvo

Zizelmann, R. (2023). Neue Formate und Onlineberatung. *Hessische Blätter für Volksbildung, 73*(4), 91–102.

19 What's Wrong with AI as a Human Rights Teacher for Social Workers? Reflections on a Teaching Innovation Project in Spain

María Inés Martínez Herrero, Francisco J. Lorenzo-Gilsanz, and Sergio Barciela Fernández

Introduction

Artificial intelligence (AI) has rapidly emerged as one of the most influential technological developments of the 21st century, permeating everyday life and reshaping professional fields from healthcare to education. Social work, a profession historically committed to human rights and ethical engagement with vulnerable populations, is increasingly exposed to the promises and risks of AI.

While the digitalisation of social services has tended to be slower and less visible than in other social policy areas (e.g. education or healthcare), the specialist *digital social work* field has been growing steadily for many decades. With the COVID-19 pandemic, however, conventional social work was suddenly 'disrupted,' and digital or e-social work shifted from being an emerging specialisation to becoming a mainstream component of practice (López-Peláez et al., 2024).

In this context of accelerated technological development, and especially after the global public release of generative AI tools in 2022, AI has experienced an unprecedented boom in visibility, accessibility, and integration into everyday practices, including the social intervention professions. This has amplified long-standing debates about the incorporation of AI into welfare and social care systems, sparking both the generally dominant optimism about its efficiency potential and the concerns about ethical and professional implications in social work (Minguijón and Serrano-Martínez, 2022).

Within the debate about the rise of AI in social work, our work starts from a cautious and critical position, as we believe that, despite AIs tools' potentialities, these cannot be understood as a neutral, harmless tools. Critical voices have emphasised how algorithmic systems reinforce inequalities, reduce professional discretion, and can underpin coercive logics within welfare systems and services (Eubanks, 2018; McQuillan, 2022; Madianou, 2025). In Spain, however, the tendency of academic and professional literature on AI and social work, following that of the digital social work field, approaches these developments with optimism, focusing on opportunities to reduce bureaucratic burdens and improve efficiency and service quality.

DOI: 10.4324/9781003627920-19
This chapter has been made available under a CC-BY-NC-ND 4.0 license.

Against this backdrop, this chapter considers to what extent an AI chatbot can contribute to critical human rights education in social work, and what pedagogical and ethical risks accompany its use. More specifically, we will present and discuss a teaching innovation project carried out at three social work courses in Spain, which we coordinated and believe demonstrates both the potential and the pitfalls of engaging with AI as a pedagogical tool for enhancing human rights in social work.

Towards Critical Human Rights Education in Social Work: Gaps, Risks, and Opportunities amid Emerging AI Tools

Human Rights in Social Work Education: A Persistent Training Gap

As emphasised by social work international organisations and professional bodies, the profession is intrinsically linked to the defence and promotion of human rights (IASSW and IFSW, 2020). In Spain, this commitment is stated in the code of ethics of the General Council of Social Work (CGTS, 2015), which requires social workers to recognise and exercise human and social rights (principle 11, p. 9), becoming

> professionally involved in the rights and interests of the service user, informing, when necessary, the competent authority and professional bodies about human rights violations, abuse, or any cruel, inhuman, or degrading treatment suffered by any person that comes to their attention in the course of their professional practice…
>
> (article 34, p. 13)

However, in Spain, as in many other countries, there remain significant educational gaps that prevent students from adequately preparing for rights-based professional practice. As research has shown, most Spanish social work degree programmes lack a strong and coherent human rights component. Only a minority of universities' curricula explicitly include substantial human rights training (Cubillos et al., 2017), and even when human rights are addressed, students tend to perceive the teaching as abstract and disconnected from real practice (Martínez-Herrero and Charnley, 2021). Martínez-Herrero et al. (2024) recently concluded, in the light of interviews with practitioners in highly sensitive human rights practice contexts in the country (migration, homelessness and older people care homes), that superficial or sporadic human rights social work education leads to frustration among social workers, who struggle to apply rights-based approaches despite being aware of violations in their daily work.

This underlines the urgency of fostering a more critical engagement with human rights within social work education. As Garrett (2024) warns, the human rights discourse in social work risks becoming superficial buzzword' if not examined through a structural, historical, and political lenses, while Ife (2016) reminds that human rights frameworks in social work can only remain relevant and fulfil their liberatory potential if applied critically rather than confined to narrow, legalistic, or individualist understandings. Without such critical reflexivity and social justice

orientation, human rights may fail to remain a relevant framework for confronting oppression and injustice in contemporary practice, and social work education risks perpetuating partial, abstract, and disconnected approaches that frustrate rather than empower future professionals.

The need to address this gap in social work education underscores the importance of developing teaching resources to help generalise such kind of critical human rights education for social work students and professionals.

This teaching innovation project must be situated within this context: a landscape where students lack coherent rights-based training, supervision, and institutional support. The starting point was that AI cannot substitute for these systemic deficiencies, but it may help make them visible and provide scaffolding for reflective learning.

A Human Rights Education Gap That Can Be Addressed with AI's Tools?

Facing the human rights educational gap in social work, the generative AI boom, and our university's funding opportunity for teaching innovation projects involving the development of innovative digital materials to enhance course subjects, we started to consider with a group of human rights social work researchers of several universities in the country whether an AI-enhanced tool could be used to realise and promote the much-needed deep and critical human rights social work education capable of transforming practice. While the pedagogical design felt promising, much attention was paid in our discussions to critical perspectives which demanded that we questioned the broader implications of adopting AI in social work education.

Evidence that emerged over the last decade points to the fact that, despite their educational contributions to professional education, AI tools such as 'teacherbots', AI chats or AI-powered case simulation tools can and should not substitute human teachers and mentors (Bayne, 2015; Segal, 2025). Furthermore, international literature has exposed how digital technologies can reinforce inequalities and undermine professional autonomy and welfare rights.

Eubanks' book *Automating Inequality* (2018) has documented how high-tech tools used in welfare systems in the United States targeted and penalised the poor with the approval of business and political stakeholders. More broadly, but also relevantly, Madianou (2025) warns against the global dangers of 'technocolonialism', showing how supposedly benevolent technologies can perpetuate harmful forms of colonialism and control. McQuillan (2022) more directly frames AI as inherently tied to authoritarian tendencies unless actively resisted, positioning critical engagement as essential. Within social work specifically, there is a growing recognition that neoliberal reforms alongside algorithmic decision-making have eroded professional discretion, creating deep frustration among practitioners (Carey and Singh, 2025). These and other critical AI authors' arguments resonate with the concerns and debates of AI implications for social work, which are increasingly recognised as central to the profession.

It is therefore essential that the broader expertise of social workers and social work service users is mobilised to critically interrogate the implications of emerging technological developments, particularly the 'disruption' generated by the rapid rise of generative AI. Social work stakeholders must become active participants in their design, regulation and evaluation, and social workers have a key role to play in 'opening AI black boxes' and influencing technological trajectories with professional knowledge and ethics.

The Teaching Innovation Project: Design and Results

Our project, *Teaching Innovation for the Integration of Human Rights in Social Work Practice Placements*, sought to address the gaps highlighted above. Led and funded by the National University of Distance Education (UNED), the project team comprised 13 recognised research experts and educators specialising in human rights in social work from five Spanish universities[1] (the learning experience for students was implemented in three of these). The team's expertise and commitment to the topic ensured the quality and depth of the project´s foundations and outcomes (AI tool, website, dissemination events, etc.). It also fostered rich, dynamic discussions and mutual learning on the topic.

The specific project aims were to:

1 **Digitalise and AI-enhance the human rights assessment tool** developed by Martínez-Herrero et al. (2024), transforming it into an interactive chatbot.
2 **Pilot the tool** with undergraduate social work students across three public universities in Spain, gathering both pedagogical feedback and research data about human rights social work education.
3 **Create a dedicated website** on social work and human rights, hosting resources, case studies, and a forum for discussion.
4 **Analyse student responses** and interactions to inform both pedagogical practice and broader research on human rights in social work.

The design of the project reflects an attempt to combine pedagogical experimentation with rigorous ethical safeguards. Specifically, the project was structured and unfolded in four stages, drawing on inter-university collaboration and collaboration with an AI development company.

Phase 1 – Design and development, focused on adapting the human rights assessment tool by Martínez-Herrero et al. (2024) into a digital format suitable for chatbot interaction and on developing a website to host the chat (now closed) and share human rights resources. The project team worked in collaboration with technical experts to ensure both pedagogical coherence and technological reliability.

Phase 2 – Dissemination and launch, the project and its digital platforms were disseminated through multiple channels to the targeted students including a public webinar for the chatbot launch.

Phase 3 – Pilot implementation, took place between April and May 2025 and involved 236 final-year students across three Spanish universities. Students engaged with the chatbot during their practice placements.

Phase 4 – Data analysis, adopted a mixed-methods approach. Quantitative data were drawn from student ratings, while qualitative insights emerged from the textual responses to the open questions. These were analysed thematically in two rounds: an initial AI-assisted analysis carried out by the developers, and a subsequent manual thematic analysis conducted by the authors. In addition, an AI SWOT-CAME analysis (Strengths, Weaknesses, Opportunities, Threats – Correct, Adapt, Maintain, Explore) synthesised findings and informed recommendations to enhance human rights social work education.

Regarding ethical considerations, the project emphasised voluntariness, transparency, and critical feedback. Students were informed about the experimental nature of the tool and its goal to generate human rights knowledge, stimulate reflection, and identify systemic gaps in human rights training. Adherence to research ethics protocols of each university ensured compliance with research ethics, including data protection and respect for student privacy.

Table 19.1 summarises the development instructions provided to the outsourced AI service for the Human Rights Chat.

Table 19.1 Summary of AI chat development instructions

	Purpose	*Key requirements*
Chat introduction and consent	Present tool purpose and obtain informed consent.	Greeting, explain aim ("formative and reflective conversation about the incorporation of HR into practice"), estimate duration (~30 min), include consent statement before continuing.
Conversation flow and time management	Structure and pace the dialogue.	15 main questions with possible follow-ups; allow relevant comments; keep within 30 minutes
Content domains	Cover three areas of the adapted HR assessment tool Martínez-Herrero et al. (2024).	Ask about placement organisation and service user profile. Then address (1) individual practice context, (2) organisational context, and (3) broader professional context.
Feedback approach	Provide formative, constructive responses.	Begin concisely and offer "read more…" option; reinforce rights-aligned answers; acknowledge barriers and suggest actions or resources for improvement.
Reflective and empowering learning	Encourage critical reflection and continuous learning.	Invite exploration of theory and models for HR practice; promote self-assessment and reflective practices, stimulate discussion and supervision.

(*Continued*)

Table 19.1 (Continued)

	Purpose	*Key requirements*
Resource integration	Direct students to authoritative materials.	Suggest using external HR (e.g. UN, Amnesty International) and social work resources, including legislation, ethics codes, and training opportunities (webinars, courses, etc.) by key organisations.
Organisational context Guidance	Frame how to discuss institutional settings.	Recognise tensions between official missions and everyday practice; promote evaluation of empowerment within organisations.
Professional awareness and Collective Action	Situate practice within wider professional and advocacy networks.	Encourage awareness of international and national social work organisations; highlight the professionals' role as HR defenders as per UN Declarations; stress collective activism's role, give examples.
Closing and evaluation	Conclude dialogue and collect feedback.	Ask students to rate (1–10), list positive aspects and suggestions for improvement, thank-you message.

Overview of the AI Chatbot Project Outcomes

Table 19.2 provides a visual summary of the project outcomes before moving into a deeper discussion of selected findings.

Overall, the project results point to the potential of AI as an educational resource to enhance human rights teaching in social work education while highlighting the need for structural curricula reform and institutional investment in supervision. Moreover, these outcomes also lead to several deeper questions such as *what exactly is missing from current human rights education in social work, and what specific pedagogical or technical improvements could strengthen AI's role as a human rights learning tool?* The following discussion explores these questions, organised around two main areas: (1) findings relating to human rights social work education, and (2) findings concerning the AI chatbot's pedagogical and technical performance.

Discussion and Critical Reflections: Human Rights Education in Social Work and AI as Human Rights Teacher

The project largely underscored the potential of using a conversational AI tool as a process of contrast for professional practice. Nevertheless, to obtain satisfactory results, certain aspects related to university education in Social Work concerning human rights, as well as the appropriate configuration of the tool for this purpose, must be considered.

The Need for Addressing Insufficient Specific Training in Human Rights and Poor Transversal Integration in Social Work Courses

A first issue emerging from the findings of a thematic analysis of the chats is the limited specific training in human rights within the social work curriculum, as well as its insufficient transversal integration in other subjects.

Table 19.2 Overview of AI chatbot project outcomes

Dimension	*Main findings*	*Notes/examples*
Overall reception	Students rated the AI chatbot learning experience highly, 8.33/10, highlighting its usefulness and innovative design.	The majority described it as "engaging," "clear," and "helpful for reflection."
Impact on reflection and learning	The personalised feedback helped students link practice situations with human rights principles, encouraging ethical awareness and self-reflection.	"It encouraged me to think about ethical dimensions I hadn't considered before."
Educational and institutional gaps revealed through thematic analysis (AI – outsourced and manual)	Thematic analysis of chat responses (AI-assisted and manual) exposed systemic weaknesses and recurring themes across all three universities: • Weak training foundations: students felt underprepared to address rights in practice. • Supervision gaps: few had formal spaces to discuss human rights dilemmas. • Fragmented resources: inconsistent institutional protocols increased vulnerability. • Commitment and interest: despite these limitations, students showed strong motivation to learn and apply rights-based approaches.	"[Was introduced to human rights based social work] only in professional practice; but I have missed specific training in this area throughout the four years of my degree."
SWOT-CAME analysis (AI-outsourced)	Strengths: Normative commitment, student motivation. Weaknesses: Scarcity of supervision, limited human rights training. Opportunities: Use of innovative technologies, inter-university collaboration. Threats: Precarious work conditions, institutional fragmentation.	These results informed recommendations for curriculum design, supervision, and ethics of using AI in social work education.
Overall interpretation	AI chatbots have strong potential as reflective learning tools, but they cannot replace structural curricular reform or human supervision.	The next section analyses these aspects in detail.

It is common to find only one optional course or isolated modules dedicated to this topic, which leads most students to complete their studies without a solid mastery of human rights–based approaches to social work practice. Examples of students' replies to questions about how they had been introduced to human rights–based social work are:

> In several subjects, models have been explained; but there is no specific one that works from the Human Rights Approach; therefore, the training in this regard is scarce and scattered.
>
> ST (student)

> The onl˚y course we have had in the degree was International Protection of Rights; but it is an optional subject; which means you might not take it.
>
> ST

This situation reinforces earlier research pointing to the lack of consistent human rights training in Spanish universities (Cubillos-Vega et al., 2017; Martínez-Herrero and Charnley, 2021) and the risk that human rights become merely rhetorical in social work practice, reinforcing the status quo (Garrett, 2024; Ife, 2016).

In this sense, the tool's training process of the tool gains becomes strategically important, as it identifies learning gaps and drives the progressive adaptation of its content and functions towards continuous improvement and the consolidation of key competencies in human rights and social work.

> [The IA chat] has provided me with information and perspectives on the lack of training in Human Rights that I have received during my university studies. It would be important to create a mandatory course specifically on this topic.
>
> ST

Lack of Formal Spaces for Supervision and Reflection

The chats also provided evidence of a notable absence of institutionalised spaces for ethical supervision and reflection during students' practical training. Supervision in social work, understood as the accompaniment of an experienced professional who helps analyse interventions, emotions, and decisions, is a key element for deep learning. However, many academic programmes in Spain do not offer this component systematically (one of the participating universities social work course does). Furthermore, most placement organisations lack a formal structure where students can discuss complex situations or human rights dilemmas in a guided manner.

> There is no [formal] supervision space; what I consider very relevant and that would be very interesting to include.
>
> ST

This finding is particularly significant in those fields most affected by human rights issues, such as migration, housing, or poverty, where social workers' interventions can have substantial consequences for the lives and dignity of the people they work with. Without spaces for rights-based discussion, students may normalise inadequate practices or overlook ethical doubts, losing the opportunity to refine their approaches or deepen their understanding of human rights principles.

Limited Coordination of Resources and Scarce Institutional Support

The chats also revealed limited coordination among social intervention services, alongside an insufficient institutional support from practice placement organisations, which together exacerbate the vulnerability of students and newly qualified

professionals when addressing human rights cases. Without single reference points, the fragmentation of information and the absence of coordination leave social workers, especially the most inexperienced, operating in isolation (Martínez-Herrero et al., 2024).

> There is no single reference to turn to; each organisation has its own protocols in this regard. This fragmentation and dispersion cause inconsistencies and different actions by each social worker in the face of the same problem; therefore, we cannot consider ourselves protected.
>
> ST

This lack of clarity about ethical action pathways intensifies professional vulnerability as social workers frequently confront high psychosocial risk situations that may also entail personal or legal dilemmas. A limited coordination ultimately undermines the quality of intervention and may compromise the effective safeguarding of service users' rights, particularly when responsibilities are unclear and support resources are unavailable at critical moments.

All these gaps support the urgent need to propose scalable solutions to expand and strengthen human rights education in social work. Taking advantage of the opportunities offered by new technological tools (notably generative AI) appears an important route to explore.

AI Chatbot Performance

While students' evaluations of the AI chat-based learning experience were highly positive, and we agree that it enhanced human rights awareness regarding social work practice contexts, contrasting the chatbot's performance with the project team's design aspirations (see Table 19 1) reveals aspects for consolidation, alongside shortfalls and areas for further development or improvement. We highlight and discuss some of these in the following subsections, distinguishing between pedagogical and technical dimensions. As social work educators, our expertise is situated primarily in the pedagogical sphere (including professional ethics), and our observations on technical matters are made from this position. A more in-depth exploration of those aspects would require interprofessional collaboration with AI and educational technology specialists.

> Using the chat to learn about human rights in social work has been a useful tool because it allows quick access to clear and structured information. It has helped me reflect on different topics and to formulate responses with a more well-founded approach.
>
> ST

> It helps reflect on knowledge; but I would prefer to share my experiences with people. The chat repeats what I say, and it's not the same as what a colleague would say.
>
> ST

Pedagogical and Ethical Design

Evaluating the pedagogical design of the tool is the first step toward substantial improvement of our tool and any other AI-enhanced tools addressed at teaching human rights to social work students and practitioners. Our experience underscores the need for an integrative and ethically grounded educational approach, where human rights content is clearly articulated. It is crucial that the tool is designed according to ethical principles, including respect for the privacy of experiences shared by students, absence of discriminatory bias in generated responses, and transparency regarding its aims and limitations. Moreover, our experience echoes the argument that AI in social work education must be guided by critical, participatory principles rather than by a technocratic logic (Boetto, 2025; Raya-Díez, 2023; Šumskienė et al., 2025).

It is also advisable to personalise the learning experience: students have different levels of knowledge and experiences, so an AI tool should adapt its questions or advice based on the user's profile. This could be achieved through initial self-assessment questionnaires or algorithms that adjust the difficulty and depth of questions according to previous answers.

Our tool included these kinds of questions and provided adjusted replies to the different answers to the specific questions.

> To what extent do you know your rights as described in specific legislation, human rights treaties and instruments, and the Social Work Code of Ethics?.
>
> AI (AI chatbot)

However, a holistic adaptation of the chatbot interaction to the user profile arises as an opportunity for further improvement and development of the AI tool. Furthermore, we detected a significant number of chats, over half of them, in which students responded with very brief and superficial (sometimes monosyllabic) answers to some or most questions. AI human rights teaching tools should be trained to detect rushed, ambiguous, or imprecise answers. In such cases, it should activate mechanisms of follow-up questioning or reformulation to ensure useful and meaningful responses. It remains an open question, and a topic for future evaluation and research, to what extent these improvements to the chatbot would reduce superficial engagement, or whether such behaviour reflects a more intrinsic tendency of human interactions with machines. Finally, consistent with critical perspectives on human rights social work education (i.e. Ife, 2016), we made sure to include reflective prompts that invite professionals to promote human rights awareness among service users and identify biases, prejudices, or paternalistic practices that contravene human rights–based practice.

> Have you considered how social work practice can contribute to the reproduction of injustices, biases, and human rights violations?
>
> AI

> Do you use specific strategies to work on rights awareness with users, helping them identify and understand these violations?
>
> AI

We consider this empowerment- and participation-oriented function a necessary feature to maintain and even develop further in AI tools aimed at teaching human rights–based social work practice.

Practical Functionality

Beyond its ethical and pedagogical coherence, the chatbot's practical functionality determines whether it can support meaningful student learning and professional reflection. For the contrast process to be effective, the tool must be both useful and relevant to real-life social intervention contexts.

One concrete recommendation is to enhance its interactivity and depth. If a student provides very brief or superficial answers, the tool should be programmed to detect this lack of development and respond accordingly. The goal is to prevent unproductive exchanges where the system merely accepts minimal input; instead, it should act as a facilitator that encourages students to expand their reasoning and explore different dimensions of a problem.

> I cannot express everything I want in the chat because the length is limited.
>
> ST

Additionally, incorporating a logical validation mechanism for students' responses could further strengthen the tool's pedagogical impact. For example, by verifying whether the answer addresses the question coherently and in sufficient detail, and if not, issuing a subtle prompt suggesting elaboration.

Another practical enhancement would be the inclusion of interactive, case-based simulations within the training process. Beyond theoretical Q&A, the tool could present dilemmas based on frequent rights violations (e.g., discrimination in access to social services, or confidentiality vs. child safety). The student would make decisions within the simulation, and the AI would provide analytical feedback on the human rights implications of their choices.

> Access to adequate housing, access to essential resources, and treatment for addictions are fundamental rights that must be promoted and guaranteed to achieve a dignified life.
>
> ST

This form of situational learning would help students practise detecting and managing rights violations in a safe environment before facing them in professional contexts. Furthermore, the tool should add substantive value to the conversation rather than merely paraphrasing the student's input. Some participants noted that the chat sometimes repeated their own words without offering new perspectives. Grounding discussions in human rights instruments, current policies, ethical codes and institutional protocols becomes fundamental.

> It is excellent that the service protocols where you work are aligned with the Social Work Code of Ethics and human-rights principles. Acting within

> a context that respects the Constitution and human rights provides a solid framework to ensure respect and promotion of users' rights.
>
> AI

Nevertheless, we found that the chatbot's references to these tended to be of an abstract nature, a limitation in line with critics of AI's tendency to context-blindness and superficiality (McQuillan, 2022). Linking such normative frameworks with case examples, drawn either from students' own placements of from simulated scenarios posed by the AI tool, becomes a fundamental enhancement needed to bridge human rights principles with practice realities.

> I find it [the AI chat] very interesting; I even think there should be an anonymous chat where people could ask questions and discuss ethical dilemmas. In my opinion, a chat is more objective and less biased than the human mind.
>
> ST

Promoting interdisciplinary coordination, teamwork, and collective action as strategies for human rights–based social intervention is equally essential. The chat performed well in encouraging the first two.

> It is very positive that you work in a proactive and coordinated team; interdisciplinary collaboration is key to effectively addressing human-rights violations. Open communication and the exchange of perspectives can enrich intervention strategies and ensure a more holistic response to service users' needs.
>
> AI

However, despite our signposting to collective action and other collective ways of advancing and defending human rights in the indications about the chatbot intended responses during the chat design and initial pilots (within the team) an important gap remained regarding the chat´s references to the latter (activism, campaigning, etc.). This led us to reflect about the possible individualistic bias of the information managed by the AI tool and the need to ensure it is also provided with materials reflecting more diverse, collective and political views of human rights–based practice, to reflect social work's broader knowledge base and social justice principles as they link with human rights.

Supported and Collective Learning, Professional Care, and Lifelong Professional Development

While no technological tool can replace the human dimension in developing professional values and judgement, it can, and should, recommend collegial and interdisciplinary discussion as an essential practice of contrast and learning.

> It is important that you use official channels to manage situations involving human rights violations; this ensures that concerns are treated with the

> seriousness they deserve. Although it may be challenging to communicate with the supervision team, it is essential to persist in seeking support, as it can offer guidance and backing in complex situations.
>
> AI

> It is also important to be able to share your concerns with colleagues, as this can generate mutual support and new perspectives. It would be useful to find out whether your university offers any supervision space to address these issues; it could provide additional guidance.
>
> AI

Supervision and mentoring are also closely tied to professional well-being. Professional self-care, often underestimated in social work education, is critical given the emotional strain and stress of dealing with complex social issues.

> No, social workers promote for others a dignified and independent life, work-life balance, and many rights; but in practice, we work long hours and must care for ourselves because we are exposed and not adequately protected.
>
> ST

This need for self-care should not be reduced to individual coping strategies alone. It must be supported by institutional mechanisms that guarantee supervision, protection, and occupational well-being. Literature on social workers' experiences of promoting human rights, including is such adverse contexts as war-torn Palestine, shows that being part of a professional community and collective perspectives to care are essential for resilience and protection in human rights–guided practice (Pérez-Viejo et al., forthcoming).

> I don't think there is enough protection for Social Workers. Although ethical standards exist, there are sometimes insufficient resources and real support to defend ourselves when things get complicated. It would be good to have more support spaces.
>
> ST

Linked to these elements, the tool should encourage continuous professional development and periodic updates. For sustaining a human rights perspective, lifelong learning and active engagement with professional networks is key. Students should become familiar with leading organisations, national and international, and understand how they promote social work practice from a human rights perspective.

> Have you received training regarding how international social work associations (for example, IFSW, IASSW, ICSW, ESWRA) promote social-work practice from a human-rights perspective?
>
> AI

A practical recommendation would be to integrate within the AI tool a digital library and regular updates on new cases, reports, and international developments, so that it can act as a channel for ongoing professional learning.

> It is essential that organisations such as the General Council of Social Work and the Spanish University Association of Social Work promote a human-rights-based practice through continuous training….
>
> AI

> Addressing these aspects through continuous training, self-care, and occupational well-being is essential to minimise risks and promote more ethical and conscious practice.
>
> AI

Supervision, self-care, collective action and reflection, and lifelong learning are not separate pedagogical dimensions but interconnected conditions for sustainable, ethically grounded professional practice. Embedding these in AI-supported learning, through prompts for reflection, recommendations, and links to resources could help translate individual reflection into collective, institutional and professional growth regarding human rights–based practice.

Some Final Technical Recommendations for the Use of AI in Training

Provided all the above, we envisage that the tool can be trained as a complementary resource for practising ethical decision-making in everyday dilemmas. Since students value the possibility of consulting the chat about complex situations, the system could incorporate a specific virtual ethical adviser mode.

> Could I use the chat if an ethical dilemma arises at some point?
>
> ST

> Of course, you can return to this chat whenever you need. If you face an ethical dilemma or any other issue related to your social work practice, I will be here to help you reflect and explore possible approaches.
>
> AI

However, as discussed, the AI should not limit itself to generic or surface-level responses; rather, it must also discern when to probe more deeply into received answers. From a programming perspective, this means equipping the system with criteria to identify ambiguous, contradictory, or overly brief responses and to activate follow-up or clarification routines.

It is equally important for the AI to recognise its own limits: if a student raises an issue beyond its training domain or requiring sensitive professional judgement, the system must be transparent about these limitations and redirect the user towards

human supervision. Ensuring such transparency about how responses are generated and what the tool's scope entails will foster trust and informed use among students.

> I understand that the current format may not have added significantly to your prior knowledge; but I am glad you were able to identify possible paths to expand your understanding on your own. Feedback like yours is valuable, as it helps improve the tool to make it more useful for others.
>
> AI

Lastly, we suggest including a synthesis or evaluative function at the end of each interaction. After several exchanges, the tool could offer the student a brief feedback report highlighting strengths and weaknesses in their responses, for instance, noting if they tended to provide concise answers, demonstrated coherence in reasoning, or overlooked important analytical dimensions.

> [Using the IA chat] has been a good experience; but you should make me a report on human rights in social work so that I would have a reference to help me carry out my work correctly. It would be a very good working tool. Thank you very much.
>
> ST

Conclusion: Promises and Contradictions of AI as a Human Rights Social Work Teaching Tool

This chapter has examined the potentials and limitations of using AI as a pedagogical tool for human rights in social work education. The teaching innovation project, involving 13 social work educators and 236 students, demonstrated that an AI chatbot can stimulate some reflection, provide personalised feedback, and expose gaps in training.

Students evaluated the chatbot learning experience very positively, giving it an overall score of 8.33/10, and they particularly appreciated its drawing attention to the topic's importance and prompting reflection on their social work practice. However, this high level of student satisfaction contrasted with a prevalence of superficial answers, suggesting limited engagement with the chatbot discussion, especially when compared with the depth of engagement typically observed in face-to-face conversations between a human lecturer and a student. The reasons for this discrepancy, as well as the broader possibilities and limitations of AI chatbots in achieving meaningful student engagement, need to be further researched.

At the same time, results revealed structural deficiencies in human rights education in the participating courses which resonate with previous data on human rights social work education in Spain: from insufficient curricular presence to limited supervision and institutional fragmentation at the practice placement organisations that no AI tool can resolve on its own. Critical literature on AI reminds us that its uncritical adoption, including as 'educator', risks deepening inequalities, eroding autonomy, and reinforcing coercive logics.

Experiences similar to our teaching innovation project offer unique opportunities by providing students with interactive, reflective tools and accessible resources. These can enhance motivation, autonomy, and rights-based awareness, as well as contribute to the development of professional networks and communities of practice. However, several noteworthy contradictions and challenges remain, which can be synthesised in the following disjunctives:

- **AI as educator vs. as bureaucrat**: While the chatbot aims to guide reflection, there is a risk of replicating bureaucratic logics of assessment, reducing rich human rights dilemmas to checklists and individualising and depoliticising deeply structural and political circumstances and breaches of rights.
- **Reflexivity vs. superficiality**: Students may provide minimal answers if not sufficiently engaged, and AI responses may reinforce superficiality rather than challenge it.
- **Empowerment vs. dependency**: Students may become reliant on AI feedback instead of developing their own critical judgement.
- **Transparency vs. opacity**: Without clear communication of the chatbot's limitations, students may overestimate its authority, potentially normalising flawed or biased responses. AI's intrinsic black boxes and social work educators' limited knowledge of AI systems contribute to opacity and reduced accountability for the educational processes.
- **Complement vs. replacement**: The danger lies in using AI as a substitute for human rights expert educators, supervision, mentorship, or institutional investment in human rights education.

What's wrong with AI as a human rights teacher in social work? is therefore not a rhetorical question. The risks are real and include the reduction of complex rights dilemmas into simplistic interactions, erosion of supervision, reinforcement of bureaucratic pressures, and the reproduction of systemic biases (such individualised approaches to human rights and professionalism). Indeed, AI should never be understood as a teacher in its own right but as a tool that, if carefully designed and ethically deployed, can complement human expert teaching, enrich student learning, and contribute to the longstanding struggle to make human rights real in social work practice. Ultimately, this experience highlights that AI (in our case a 'teacherbot'), like an increasingly sophisticated brush, can only achieve meaningful educational outcomes when guided by the ethical judgement, critical awareness, and contextual understanding of the human hand that directs it. As Raya-Díez (2023) notes, AI can contribute ethically if placed at the service of people and grounded in professional values, but this requires vigilance, reflexivity, expert input and collaboration and institutional accountability.

AI Statement

During manuscript preparation, the authors made limited and selective use of AI (ChatGPT, version 5.2, by OpenAI) to assist with translation and proofreading. All

content was human-led and revised by the authors, who retain full responsibility for the accuracy, interpretation, and final text of the chapter.

Note

1 UNED, University of la Rioja, Complutense University of Madrid, Pontifical University of Comillas, and University of Castilla la Mancha.

References

Bayne, S. (2015). Teacherbot: Interventions in automated teaching. *Teaching in Higher Education*, *20*(4), 455–467. https://doi.org/10.1080/13562517.2015.1020783

Boetto, H. (2025). Artificial Intelligence in Social Work: An EPIC Model for Practice. *Australian Social Work*, 1–14. https://doi.org/10.1080/0312407X.2025.2488345

Carey, M., & Singh, G. (2025). Introduction to the special issue: Social work and social control. *Critical and Radical Social Work*, *13*(1), 2–6 https://doi.org/10.1332/20498608Y2024D000000058

Consejo General del Trabajo Social (CGTS) (2015). *Código deontológico del trabajo social en España*. Consejo General.

Cubillos-Vega, C., Ferrán Aranaz, M., Mercado García, E., & Pastor Seller, E. (2017). Nociones y compromiso con los derechos humanos en estudiantes de Trabajo Social en España. *Revista Mexicana de Investigación Educativa, 22*(75), 1047–1075. ISSN 1405–6666

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Garrett, P. M. (2024). Human rights and social work: Making the case for Human Rights Plus (HR+). *The British Journal of Social Work, 54*(5), 2107–2123. https://doi.org/10.1093/bjsw/bcae022

Ife, J. (2016). Human rights and social work: Beyond conservative law. *Journal of Human Rights and Social Work, 1*(1), 3–8. https://doi.org/10.1007/s41134-016-0001-4

IASSW and IFSW (2020). Global standards for social work education and training. https://www.iassw-aiets.org/global-standards-for-social-work-education-and-training/

López-Peláez, A., Kalenda Vávrová, S., Gojová, A., Erro-Garcés, A., & Acebes Valentín, R. (2024). Digital social work. *International Social Work*, *68*(4), 728–734. https://doi.org/10.1177/00208728241267804

Madianou, M. (2025). *Technocolonialism: When technology for good is harmful*. Polity Press.

Martínez-Herrero, M. I., & Charnley, H. (2021). Resisting neoliberalism in social work education: Learning, teaching, and performing human rights and social justice in England and Spain. *Social Work Education*, *40*(1), 44–57. https://doi.org/10.1080/02615479.2020.1747421

Martínez-Herrero, M. I., Lorenzo Gilsanz, F. J., & Barciela Fernández, S. (2024). Seeking to fulfil the human rights promise: Practitioner perspectives and an assessment tool for realizing human rights based social work education and practice. *Social Work Education*, 1–18. https://doi.org/10.1080/02615479.2024.2416571

McQuillan, D. (2022). *Resisting AI: An anti-fascist approach to artificial intelligence*. Bristol University Press.

Minguijón, J., & Serrano-Martínez, C. (2022). La inteligencia artificial en los servicios sociales: Estado de la cuestión y posibles desarrollos futuros. *Cuadernos de Trabajo Social, 35*(2), 319–329. https://doi.org/10.5209/cuts.78747

Raya Díez, E. (2023). Artificial intelligence and social work: Contributions to an ethical AI at the service of people. In A. López-Peláez & G. Kirwan (Eds.), *The Routledge international handbook of digital social work* (pp. 368–381). Routledge.

Segal, M. (2025). Social workers' evaluation of ChatGPT for solving ethical dilemmas within the limits of confidentiality. *Journal of Social Work Practice*, 1–14. https://doi.org/10.1080/02650533.2025.2480092

Šumskienė, E., Bald, C., & Martínez M.I. (2025) "Chapter 7: Even if Elon Musk was a social worker…": Coercive past and technological futures in social work in Lithuania, UK and Spain. In G. Ottmann & C. Noble (Eds.), *AI and the disruption of welfare challenges for social work education and practice*. Routledge.

Index

Note: **Bold** page numbers refer to tables.

Aavik, K. 118
ableism/ableist 206–207; bias 207, 212; conceptualizations of 206; hierarchies 207; racialized 206
accountability 9, 73, 95, 103, 135, 149, 157–158, 162, 164, 171, 174, 229
action, politics of 68
Adams, S. 77
advocacy 2, 8, 14, 20, 24, 42, 68–70, 153–154
affordability 30, 62, 66–67
Age of Surveillance Capitalism, The (Zuboff) 16
Agger, B. 95
Aginskaitė, S. 11
Ahn, E. 148
AI-based counselling in social work 233; expert discourse on 234–236; implications of 236–237; potential impact of 236–238
Alamdar-Niemann, M. 11
algorithmic: bias 144, 153, 156–157, 219; decision-making 247; epistemic injustice 117; governance 228–229; opacity 163, 228; systems 18–19, 154, 174, 219
Anthropic 1, 7, 8, 20
Ardalan, I. D. 104
Aris, S. 9
Aristotle 74, 142
Arnason, J. 81
artificial emotional intelligence 8, 50
artificial general intelligence (AGI) 49, 72, 169
artificial intelligence (AI) 1, 10, 27, 72, 99, 104–105, 107–108, 112, 130, 139, 153, 161, 178, 206, 219–220, 240–241; alternative governance of 119–120; as autonomous, human-like entities 15–16; caring professions in age of AI 145–146; chatbot 9, 11, 126–127, 130–131, 133, 135, 145, 180, 234, 246, 250–252, **251**, 253–254, 259; chat development instructions **249–250**; client communication and documentation of 156; computation and superintelligence 48–49; and cultural security 104–105; definition of 169; deployment of 9; design 6, 211; development of 3, 8, 11; digital systems 5; disparities and justice-related issues 146; effect on environment 146; emerging pedagogical dilemmas 222–223; enabled learning 7; enhanced future of care 147–148; epistemic bias in 18; ethics 118, 148, 173; futures 179–181, 186; humans and possible behaviour of 47–48; implementation 160; incremental potentials of 221–222, 224, 226–228; and Indigenous knowledge 101–104; integration 220; integrity and adoption 92; knowledge products 2; lifecycle 161; limitations of 10; literacy and professional proficiency 11; monitoring and evaluation 52–53; and politics of care 147; "positive" evaluations of 158; problems 46–48; risks 1–2; securitisation 105–107; self-awareness and self-improvement 49; in service of care 142; social construction of 94; social dimension of 94; social impact of 9; in social work 4, 10, 142, 148, 154–156, 186, 220, 223–224, 227, 229–230, 235, 247, 254;

and society 100–101; specialised ethical 170–171; student experience 221; SWOT-CAME analysis 249; teaching experience 221; technologies 6, 11–12, 60, 63, 75, 91–92, 99, 112, 114–115, 141–142, 144–145, 148–149, 181, 208, 214, 219, 222, 230; tools 10, 60–61, 63, 66, 68, 87, 92–93, 143, 145–149, 161, 213, 225, 229, 246, 255; in training 258–259; trans and nonbinary exclusion in 115–116
artificial superintelligence 46
Asian Pacific Association Social Work Educators (APASWE) 69
assistive technologies 28, 62, 69, 207–209, 211, 214, 216
Astleitner, H. 8
Astray, A. 11
augmentation 28, 33–34, 42, 148, 158
Australia Association for Social Workers (AASW) 69
authenticity 133–134, 199
authoritarian/authoritarianism 2, 18–19, 78, 146, 247
automated/automation 157–158; decision-making (ADM) 87; systems 133, 147, 174, 213, 224; transcription 158, 223, 226–227
autonomy 11, 79–81, 115, 160, 196–197, 260; Castoriadis project of 80; critical 81; in family-oriented systems 197

Bailey, D. E. 117
Balkow, M. 10
Ballestero, A. 10
Banks, J. A. 184–185
Barnes, M. 140
Bauman, Z. 23
Bean, M. 32–33
Bearman, M. 90
Beatty, C. 235
Behnia, B. 127–128
Bélisle-Pipon J.-C. 132
Bell, K. 144
beneficence 51, 159–160
Benossi, L. 186
bereavement care 190–192, 196, 200; autonomy in family-oriented systems 197; boundaries of representation 199–200; emotional and psychological safety 192–196; family mediation and participation 197–198; interventions 192; limitations and implications 202–203; memory 200–203; relational ethics 196–197; returning to life 203; ritual design 198–199; social justice 200–202; tacit knowledge and gestalt of grief 191; VR 192, **193–195**
Bernecker, S. 186
Berridge, C. 144, 148
Betlemidze, M. 133–134
Bezos, J. 113
bias 47, 65–66, 73, 112, 115, 144–145, 164, 171–172, 210, 222–223, 254; in AI systems 156; algorithmic 156–157; discriminatory 254; epistemic 18; individualistic 256; mitigation 164; problem of 114–115; and representation 209–210; systemic 260
big data analytics 28, 112, 220
Big Tech 89, 186–187
binary gender 115–116
Blas, Z. 112
Bogart, K. R. 207
Bolton, G. 182
Born, K. 41
Bostrom, N. 49
Boyle, J. 29
Braidotti, R. 141
Brain–Computer Interface (BCI) 15
Brest, P. 41
British Association of Social Workers (BASW) 147–148, 186
Brittan, A. 112
Brown, J. 93, 115
Bustio-Martínez, L. 224

Campana, M. 11
capacity *vs*. normative authority 171
capitalism 1, 3–5, 8, 54, 72, 75–78, 121, 141
care 143–145, 147, 149, 196, 215; aged 173; complexities and systemic dimensions of 140; continuity of 226; delivery 146; dimensions of 10, 174; enhanced future of 147–148; environmental 144; ethics 140, 170; human-centred 145; management 183; for more-than-human world 141–142; possibility of reimagining 143–144; practices 150; redefinition of 141; scope of 149; in social work 140–141; as sociopolitical praxis 140; technocratic replacement of 149; technologies of 182; traditional principles of 139; Tronto's framework of 139–140; *see also* bereavement care; *specific types*
Carrigan, M. 78, 82–83

case management 24, 220, 225, 228
Castoriadis, C. 72, 74–84
Černý, M. 224
chatbots 18, 60, 113–114, 126–127, 132–133, 135–136, 143, 145, 181, 235–241; human interactions 134; learning experience 259; natural language 221; performance 253; references 256
ChatGPT 87, 92–93, 100–101, 114, 145, 153–156, 159, 161, 163–164, 236–237; for communication in crisis 237; ethical implications 154; legal and regulatory considerations 159; limited emotional understanding 159; superficial analysis 159
Chaudhari, J. 76
Chihota, C. 10
Chihota, T. 10
Child, Youth and Family Services Act 159
child protection 154, 158, 170, 173, 182, 200
Clare, E. 217
class-based epistemologies 11, 224
co-creation 198, 211
Coeckelbergh, M. 17, 29
coercive control 9, 65
cognitive behavioural therapies (CBTs) 154, 233–235
collective: action 14, 119, 256, 258; learning 256–258; responsibility 67, 140, 149; rights 231; spaces 227
Collini, S. 88
colonialism/colonisation 3, 4, 8–9, 11, 21, 64, 103, 112, 206, 224
commodification 202, 230
communication 37, 63, 65–66, 126; in crisis 237; networks 3; processes 3; technologies 3, 61
competence 127–128, 160, 164
compliance 50–51, 155–157, 160, 162–163, 209, 229, 240, 249
computation 8, 48–49
confidentiality 145, 155–156, 164
Connelly, P. 182
Contanza-Chock, S. 115
contestability 171–172, 175
conversational agents (CAs) 235
co-optation 39–40
corporate: conglomerate 2; epistemic responsibility 7, 22; university 87–94
corporate power, concentration of 16; conventional knowledge production 16–17; epistemic participation and co-creation of knowledge 17–18
Couldry, N. 112
counterfactually augmented data (CAD) 156
Crawford, K. 114, 118
critical: analysis 6–7; autonomy 8, 78–81; human rights education 246–247; policy studies 94–95; reflection 6, 94, 219–220, 230, 236, 240, 250; reflexivity 246–247; scrutiny 221, 246; social theory 4–6, 8; social work 2, 8–9, 24, 137; theory 6, 8, 72, 74, 78, 81–82; thinking 10–12, 24, 73, 78, 146, 158, 174, 177, 181
Crockett, M. J. 17
Cropley, D. 76
cultural/culture 2, 6; expressions 101, 104; identity 2, 11, 101, 106–107, 206; narratives 207; responsiveness 236; security 100–101, 104–105, 107–108; transformation 107

Dai, Y. 116
Darcy, A. 234
data/database 33, 60, 130, 177, 182, 184–185, 228; analysis 61, 249; classification 105; collation and calculation of 183; collection 73, 145, 147, 161, 209; colonialism 112, 119; corporate control over 18; distortion 104; essentialism 224; gathering 209; protection 11, 104, 153, 156, 157, 233, 237; selection and modelling 147; summaries of 185; tools 219
dataism 111–112, 120
Data Protection Impact Assessments (DPIAs) 156
decision-making 10, 27, 83, 89, 94–95, 118, 143–144, 163, 174, 186–187, 229–231; economic 37; ethical 11; power 227–228; process 3, 76, 118, 129, 134, 169, 185, 227–231
deep learning (DL) 60
deprofessionalisation 183–184
Díaz, S. 101
digital/digitalisation: access and tools 67; assets 23; authoritarianism 2; automation 1, 12, 15; change, inevitability of 90; commons 119; communication circuits 5; divide 63–64, 67–68, 89, 230; economy 5, 22, 24; education 164; enclosure 93; environment 66–67; exclusion 8, 63, 242; gender-sexual violations 117; ghost 202; human rights 66–67; injustice 64; literacy 163, 227, 230; platforms 248; programming 28–29; skills 67; of social

services 245; socio-technical drivers of 89; surveillance infrastructure 3; technologies 142, 247; transformation 126–127; violence 117
Digital New Deal 29
D'Ignazio, C. 112
dignity 22, 163
disability 208–210, 213–217; assistive technologies in 209; discrimination 210; expertise 217; justice 11, 206; overcoming 208; public perceptions of 207; societal perceptions of 11; speech 215
discrimination 2, 116, 148, 236, 254
disenchantment 134
disinformation 38, 67, 106
dissemination 19, 27, 248
distortion 95, 103–104
Doerr, M. 224
domestic violence 117, 226, 228–229
Drage, E. 115
Dritsas, E. 224
Dumman, J. 239
Dunn, D. S. 207
Dustin, D. 184
Dyke, C. 183

eco-fascism 142
ecological systems 141–143, 148
Economic and Social Council (ECOSOC) 68
EdTech 93
Elliott, A. 181, 186
Ellis, K. 207
emancipation 39, 107–108
embodiment 9, 130
emotional: attachment 214; containment 199; escalation 196; flooding 198, 201–202; intelligence 169; pacing 192; safety 192–196; saturation 200; sensitivity 235
enchantment 133–134
engagement 8–9, 11, 38, 42–43, 63, 72, 78–79, 88–90, 93–94, 128, 130–131, 133–136, 197–198, 202, 220–221, 235, 245–246, 254, 257, 259
enhancement 39, 208, 214, 255–256
environmental care 144
epistemic: agency 7, 17, 24; alignment 7, 20–22; architecture 7, 16–17, 19; authority 19, 103, 171; autonomy 8; conformism 115; differences 103; dislocations 103; displacement 102; integrity 224; justice/injustice/harm 9, 19–20, 22–23, 95, 116–118; orientation 103; participation 17–18; pluralism 223; power 7; resistance 9; responsibility 7, 20, 22; risks 222; sovereignty 9, 103, 106; technology 6; vulnerability 17
epistemology 99–100, 224–225, 227
ethical/ethics 40–41, 170, 187; AI implementation 160; blue-washing 160; of care 118; challenges 213–214; decision-making 11, 258; frameworks 161; and generative AI 184–185; implementing 51; issues 148; of non-human entities 177; reasoning 230; of resistance 81; violence 118; vulnerabilities 192–193; washing 118
EU Artificial Intelligence Act (Regulation EU 2024/1689) 157–160, 210, 220, 240
Eubanks, V. 105, 247
eugenics: movement 207; technology-enabled shift in 207
EU's General Data Protection Regulation (GDPR) 157
evidence-based practice (EBP) 180
exoskeletons 208, 216
explainability 52, 158
exploitation 16, 21, 63–64, 77, 112, 115, 117, 196, 200

fairness 145, 156–157
family 201; care 202; mediation and participation 197–198; oriented systems 197
farewell 191–192
fluid certainties 177
Flynn Effect 48–49
Frabetti, F. 115

Gabriel, M. 37, 221
Garrett, P. M. 185, 246
geek masculinity 113
gender 65, 115–116; bias 18, 111, 114–115; binarism 9, 112, 115–116, 119; dimorphism 112; violence 116–118
General Council of Social Work 246
General Data Protection Regulation (GDPR) 156, 240
generative AI (GenAI) 75–79, 87–89, 117, 160, 178–179, 196, 199–200; application 73; computational paradigm of 76; ethics and 184–185; social work assessments 181–184; tools 92–93, 245; *see also* chatbots
Gilligan, C. 140
global advocacy 68–69
Global North 4, 5, 21, 61, 63–64, 113–114

Global South 3, 18, 61, 63–64, 68, 156, 236
Goldkind, L. 161–162
Gonzalez, A. 11
Great Recession 30, 34
grief 191–192, 196, 198–199, 202

Habermas, J. 3–5
Haidt, J. 179
Hall, M. 117
hallucination 199–200
Haraway, D. J. 104
harm 146, 148, 154, 174, 182, 201; to children 179; cumulative 210; relational 9
Harper, J. 94
Heinl-Vako, S. 11
Henderson, M. 90
Herrera-Semenets, V. 224
higher education 31–33, 35
Hoche, M. 144
Hodgson, D. 17
Hoffman, J. 181
holistic care 143
human: AI collaboration 161–162; AI moral symbiosis 171; bias 52–53; centred care 145; centred ethics 139; circuit breaker 191; dignity 231; intelligence 27, 48, 50–51; judgment 158; labour 5; productivity 2; reasoning 222, 224
Humanities, Arts, and Social Sciences (HASS) 88
human rights 23, 118, 246; instruments 255; perspective 257; profession 239; resilience and protection in 257; social work education 254; violations 256–257
human rights education in social work 252–253, 259; AI's tools 247–248; teaching innovation project 248–250; training gap 246–247
human services 174; ethical dilemmas in 170; trust in 174
hypermodern 7, 32, 34–35, 40–42; advanced technologies in social disruption 27–31; description of 31–34; ethical considerations 35–38; scenarios framing social work in 38–40; smart machines as social problem 27; social work strategy in 40–42; truncation of work in 34–35
hypernormalisation 32

ideology of patriarchy 112
Ife, J. 246–247
Ignatius, D. 29–30
image-based abuse 116–117
inclusive design 211, 216
inclusivity 141, 236
Indigenous: communities 65, 100, 102–104; cosmologies 105; cultures 101; epistemologies 9; inequality 65
Indigenous knowledge (IK) 99, 101
individualistic bias 256
individual-level practitioner 69–70
inequality 38, 61, 112
information: production 19; systems 33
Information and Communication Technologies (ICTs) 60, 66; and AI 60–61; deprivation implications 65–66; digital divide 61–62
informed consent 155, 164, 213
innovation 9, 14, 217; ecosystem 217; participatory models of 231; projects 247
institutional: adaptability 9; biases 154; choices 131; inequality 93; policies 227
instrumentalism 74–75
interdisciplinary audits 230
International Association of Schools of Social Work 239
International Federation of Social Workers (IFSW) 68, 148
International Non-Governmental Organisations (INGOs) 68
International Telecommunication Union (ITU) 61
Ioakimidis, V. 177

James, P. 160
Johnsen, K. I. 102
justice 146, 160, 163

Kahneman, D. 183
Kant, I. 185
Kapidakis, S. 224
Kemp, L. 14–15
Kent, M. 207
Khawaja, Z. 132
Kimmerer, R. W. 101
Klein, L. 112
Klooger, J. 80
knowledge 2–3, 33, 49, 64, 81, 102–103; co-creation of 17–18; commodification of 230; communities 99–100; creation 6; and experiences 254; holders 99; in oppressive power 18–19; production of 9, 18–20, 223; scientific 103
Kohlberg, L. 172
Kreuger, L. W. 179–180

Kuhn, T. S. 103
Kulveit, J. 100
Kumar, A. 154
Kurzweil, R. 15–17, 19, 22

La Greca, J. 221
Laming, L. 182
large language models (LLMs) 7–9, 14–15, 17–18, 22–24, 72–75, 77–79, 81–83, 87, 153, 157, 159–160, 236; in academic scholarship 73–74; critical scholarship in 81–82; design of 76; ethics of 83
leadership 93–94
learning 53, 101; pathways 220, 222; profiles 221; student 222; trajectories 222–223
Leedham, M. 179
liberal anti-discrimination discourses 114–115
Lillis, T. 179
Lindgren, S. 95, 115
Linkola-Aikio, I. 104
literacy 223, 225, 230
Little, B. 112–113
lived experience 172, 215–216
Lived-Experience Review Board 156–157
local advocacy 68–69
Lockhart, E. 113
Lu, J. H. 158
Luddite movement 36

machine learning (ML) 60, 87, 115–116, 144–145, 223–224, 228
Madianou, M. 247
Maglajlic, R. 177
management practices 143–144
Manasa, G. M. 222
marginalised communities 40, 67–68, 171–172
Marini, I. 208
Martínez-Herrero, M. I. 246, 248
masculinism 9, 112–114, 119
Matei, A. 32
McQuillan, D. 115, 117–119, 247
Mejias, U. 112
memory 27–28, 173, 200–203
mental health 155; programs 154; services 144; triage 173
mentoring 257
Messeri, L. 17
Microsoft Copilot 92
micro-targeting 2
Miragoli, M. 115
moral: development theory 172; environment 175; judgement 170; particularism 170; reasoning 170, 174; status and agency 172; value 183
morality 170, 172
Morgenstern, J. 181
Morley, C. 161
Moxley, D. P. 7–8
Müller, V. C. 49
Munhwa Broadcasting Corporation (MBC) 191–192; *I Met You* 191–192, 196–200, 202
Munro, E. 182–183
Musk, E. 113, 181

National Association of Social Workers (NASW) 147–148; AI-specific guidelines 160; Code of Ethics 155, 159; Standards for Technology in Social Work Practice 153; technology standards 159–160
National University of Distance Education (UNED) 248
natural language chatbots 221
natural language processing (NLP) 181, 223–224
Neuralink 181
neural networks 99
Nikoletos, N. 80
Noble, C. 8
Noble, S. 115
nonverbal communication 126
Nordic Network for Language Models for Artificial Intelligence 106
normalcy 206
normalisation 159, 171
normative individualism 212
norms 170
North American Free Trade Agreement (NAFTA) 32
Norvig, P. 169
Nyholm, S. 54

opacity 157, 163, 228, 260
OpenAI 92–93, 145, 153
openness 52
oppression 5, 14, 18–19, 116, 206
Ornellas, A. 183–184
Ottmann, G. 7–8
overreliance, risks of 158

Page, L. 16, 113
palliative care 196, 198–202
Papaspyridis, A. 221

Papouli, E. 9–10
Parker, S. 182
participation/participatory: design processes 230; development processes 235–236; values 41; washing 119
Parton, N. 184–185
paternalism 196–197
patriarchy 112–114
Pavani, L. 9
Pepi, M. 29
Peponakis, M. 224
Perera, M. 104
personality 20–22
personalized smart technologies 28
plagiarism 73
Plana, D. 158
Polanyi, M. 191, 197, 200–202
policy 94–95, 210–211
politics of action 68
Popper, K. R. 103
populism 30, 34, 112
post-anthropocentrism 139
post-conventional reasoning 172
post-deployment monitoring ledger 157
posthumanism 139
post-traumatic stress disorder (PTSD) 155
power 2, 199–200; colonial structures of 21; hierarchies of 1
predictive analytics 142, 153, 219
predictive care 143
pre-emptive policing 117
prejudice 212, 254
privacy 73, 144–145, 155–156
problem of bias 114–115
procedural justice 161
Professional Association of Social Work in Germany 239
professions/professionals 132, 135, 178, 180, 183, 187, 226, 231, 233, 239–241; action 129; adaptation of 33; associations 161; autonomy 226, 230–231; care 256–258; competence 158, 164; development 163; dynamics of 184; embodiment and expression 130; ethical and relational foundations of 230; ethical vigilance and epistemic resilience 220; health and caring 145–146; helping relationships 127; human rights-based 67; identity 239; interpretations 129; judgement 155, 161, 226; knowledge 180; lags 160; learning 258; practice 226, 250; principal ethical reference 159–160; principles of 186; relationships 146; training 131; universities 161; values 160; 'writing-intense' 179
psychoeducation 159, 234
psychological safety 190, 192–197
psychosocial: crises 236; vulnerabilities 192–193
public trust 77, 175, 236

Qian, Y. 100
qualitative coding 224
quality of care 144–145

race/racism 65, 112, 115–116
Ray, K. 160–161
Raya Díez, E. 260
reciprocity 9, 101–102, 131, 134–136
Redes Urbanas de Apoyo (RUA) 228
reflection 252, 258
reflexivity 126, 260
regional advocacy 68–69
relational ethics 141, 196–197; autonomy in family-oriented systems 197; family mediation and participation 197–198
relational harm 9
Ricaurte, P. 118
Richards, P. 43, 112
rights-based awareness 260
Riley, S. 224
ritual design 198–199
robotics 50–51; caregivers 215; developments in 234; laws of 50–51; systems 143
Russell, S. 169

safety 20–22, 52, 213
Said, E. 103
Salmi, J. 54
Saltmarsh, J. 94
Sámi Ownership and Data Access (SODA) principles 106
Sanfelici, M. 9
Schindler, O. 172
Schmidt, J. 237
Schon, D. A. 37
science, technology, engineering and maths (STEM) 114
securitisation theory 105–107
self-awareness 8, 49, 172
self-care 257–258
self-improvement 8, 49, 169
Selwyn, N. 90
service users 127–128
sexism/sexuality 112, 115–116
Shakespeare, W. 207

Shew, A. 217
Shipley, G. P. 2
Shklarski, L. 160–161
Shuttleworth, R. 8–9
Silicon Valley 113
Singer, P. 180–181
Singh, N. K. 154
situational learning 255
SKYLINK access 67
Small Modular nuclear Reactors (SMRs) 3
smart machines 29–32, 43
smart technologies 34, 37
social: activism 68; advancement 40; anthropology 224; care 126, 147, 156, 169, 225; change 95; crisis 4; disruption 1, 27–31; exclusion 62–63; inclusion 11, 215; intervention services 252–253; justice 24, 146, 149, 200–202, 231, 246–247; movements 145; organisations 227–228; policies 140; possibilities for 94–95; problems 177; rationalisation 5; relations 104–105; systems 5; theory 5–6; workers 225–226; work perspective 2
Social Development Goals (SGDs) 66
social work 39–41, 127–128, 132, 139–140, 143, 149, 162–164, 177, 185, 239–241; ChatGPT ethical implications 154; courses 250–252; curriculum 250–252; digital technology in 142; education 2, 220, 222, 246–247, 250; efficiency and effectiveness of 143; ethical distinctiveness 135; ethical principles in 159; ethical questions in 144; generative 178–179; generative AI 181–185; human rights education in 253; innovations 42; intervention 225; McDonaldization of 184; perspectives of 146, 180; profession 236; research 223, 225; stakeholders 248
Social Work Code of Ethics 255–256
social workers 37–38, 130, 153, 228
society 6–7, 28, 31; contemporary 77; open-endedness of 77
socio-technical: imaginary 9; systems 9, 94–95, 127; university 91
South North 5
Spain: academic programmes in 252; human rights social work education in 259; universities 252
speech disabilities 215
SPIRIT-AI 158
stakeholder inclusion 161
state surveillance 18–19, 38–39, 145, 154, 185, 213, 223, 230
statistical/computational bias 52–53
Staub-Bernasconi, S. 239
Stebnicki, M. A. 208
stigmatisation 236
structural: inequality 2, 62, 67, 95, 105, 114, 146, 221, 231; injustice 115, 163, 192, 201
students 173, 221, 240; intellectual autonomy 222, 230; learning 222, 255, 260; logical validation mechanism for 255; misuse of AI 92; practical training 252; satisfaction 259; use, industry development and regulation of 91; vulnerability of 252–253
Šumskienė, E. 11
super-crip narrative 207–208
superficial analysis 159
superficiality 159, 256, 260
superintelligence 8, 48–49
supervision 252, 257–258
supported learning 256–258
surveillance 2–4, 16, 18, 22, 38–39, 132, 145, 154, 171, 213, 223, 230
Susskind, D. 32, 94
Susskind, R. 32, 94
symbolism 198, 200
systemic bias 52–53, 65, 154, 260
systemic encroachment 3–4

tacit knowledge 191, 197
teacherbots 247
tech industry 113–114
technical: knowledge 131; literacy 235
technoableism 206–207, 212; concept of 208; narratives 212; and super-crip narrative 207–208
techno-capitalism 77–78, 80
technocolonialism 247
technological/technology 27, 32, 38–40, 53, 60, 64, 68, 104, 106, 113, 131, 139–140, 145–146, 148–149, 162, 179, 200, 202, 211–212, 214, 217, 230–231; advancements 29, 37, 236; assistive 214; barriers 67–68; of care 182; economy 30; exclusion 65; genomic 37; imperative 197; innovation 219; integration 207; personalized smart 28; physical and cognitive 30; in premodern 40; rationality 108, 113; reasoning 170; regulating 39; standards 160; transformation 89; visual features of 190–191; *see also* artificial intelligence (AI); digital/digitalisation
Teo, S. 115, 118
Thiel, P. 113

Thiessen, B. 237
Thomas, E. 117
Toli, L. 222
training: and education 161; gap 246–247; rights-based 247
trans-corporeal surrogacy 133
transcription 225
transformational creativity 77
transhumanism 113–114, 216
transparency 52, 73, 83, 157–158, 260
Trigka, M. 224
Tronto, J. C. 140–141, 149
Trump, D. 31–32, 34
trust 126–129, 131–133, 175, 236, 259; in AI chatbots 130–135; building 127, 129, 135; definition of 127; in helping relationship 127–130; in human services 174; public 77
Turner, V. W. 198, 200
Twiner, A. 179

Ullrich, C. 76
unbounded representation 201
uncertainty 129–132
UN-Habitat Digital Inclusion Playbook 66
universities 90–93, 95, 246; in age of AI 89–90; datafication of 89; discourses 90; innovations and responses 91; in neoliberal context 88; public framings of AI in education 90–91; Spanish 252
unrestrained capitalism 1
Úriz, M.-J. 10
user-centred adaptation 235–236
U.S. Health Insurance Portability and Accountability Act (HIPAA) 157
Uuk, R. 100

Valiant, L. 43
van Toorn, G. 9
Verdegem, P. 119
'ViKL' (Virtual Client) project 234
virtual assistants 143
virtual ethnography 190–192, 202
Virtual Hope Box 154–155
virtual reality (VR) 190, 201; bereavement **193–195**, 196, 198; representation in 199
virtue ethics 170
Viscarret, J. J. 10
vulnerability 9, 17, 29–30, 127–131, 135, 199–200, 202

Wach, K. 156
Wajcman, J. 119
Wallerstein, I. 4–5
Wang, Y. 235
Webb, S. 180, 182
Weizenbaum, J. 174
welfare systems 10, 105, 132, 134, 137, 140, 245, 247
well-being 39, 50, 139, 141, 143, 146, 163
Wiener, N. 40
Williams, D. H. 2
Wilson, B. 181
Wilson, W. J. 27
Winch, A. 112–113
Works Progress Administration in Depression 41

Young E. 119

Zuboff, S. 4, 16, 28
Zuckerberg, M. 113