CHAPTER 1

Introduction

Rebekah Rousi, Catharina von Koskull, and Virpi Roto

Introduction

We are now living in an era where century-long dreams of independently moving and seemingly thinking machines are becoming a reality. From robots and self-driving vehicles, to personal assistants embedded within our word processing tools, art and design generators, as well as artificial companions, all enabled by machine learning (ML). Perhaps, more aptly put, there is no sector within modern societies in which AI has not been ideated and conceptualised, developed for, and/or implemented. It is indeed becoming the 'new electricity' of the twenty-first century, empowered by the groundwork of the contemporary landscape of global connectivity and the Internet.

R. Rousi (✉) · C. von Koskull
University of Vaasa, Vaasa, Finland
e-mail: rebekah.rousi@uwasa.fi

C. von Koskull
e-mail: catharina.von.koskull@uwasa.fi

V. Roto
Aalto University, Espoo, Finland
e-mail: virpi.roto@aalto.fi

R. Rousi et al. (eds.), *Humane Autonomous Technology*,
https://doi.org/10.1007/978-3-031-66528-8_1

Despite the excitement of bountiful opportunities arising from the developments, launches of AI-based innovations have been met with scepticism, worry, and outright horror when new solutions for old problems bring with them, new problems in themselves. To illustrate, one of the attractions behind self-driving vehicles is that of reducing uncertainty induced by varying skill levels in human drivers. In other words, on the roads, some people are better at handling fast moving vehicles than others. In March 2018, belief in the integrity of self-driving vehicle safety was desecrated when a self-driving car, supervised by a 'safety driver', hit and killed a pedestrian (Elish, 2019). The physical accident itself gave rise to a number of dynamic factors that need to be considered when implementing autonomous technology such as: the human dimension—how to ensure humans maintain oversight of unfolding events; as well as accountability and responsibility—when incidents occur who should be made responsible and accountable if a machine makes decisions that are erroneous or harmful. These types of incidents, as well as other unfolding scenarios such as gender bias in AI-driven recruitment tools as well as racial bias in law enforcement technology are a part of a growing ecosystem of AI-related ethical concerns that perhaps not so much invent new dilemmas, but rather, amplify old ones.

Newer examples can be seen in the rising amount of scandals specifically related to generative artificial intelligence (GenAI) such as the *Sports Illustrated* incident in which Chief Executive Officer Ross Levinsohn was fired after using GenAI to produce stories and generated false biographies (Goldman, 2023). Another incident saw a financial official pay 25 million US dollars after engaging in a video call with a deep fake impersonating his organisation's chief financial officer (Chen & Magramo, 2024). On an increasingly personal and sensitive side of human integrity, there have been increasing accounts of deep fake porn and AI-enabled sextortian (Satter, 2023; Singer, 2024). These incidents have seen a greater number of female targets, and particularly, underage female targets (Rousay, 2023). On the flip side, ethical concerns have also caused overreactions in the systems, witnessing 'reverse biases' in output such as seen in the recent Gemini.AI case in which the AI-image generator refused to produce 'white' subjects in artistic images. This occurrence entailed visualised over-compensatory phenomena such as culturally diverse Nazi-era soldiers (see Heath, 2024). Thus, the linguistic and representational nature of our current AI-landscape is transforming tangible and physical ethical issues into cultural, sociological and psychological ones.

While the obvious catastrophes are noted and mediated, the subtler effects of how humans live with, are affected by, and in turn affect intelligent technology remain under the radar. There is no doubt that notice has been given to conditions such as the 'Google Effect', or degenerated memory (short-term, working and long-term) based on the convenience of the Google search - removing the need for memorization and thus decreasing cognitive training in humans (Sparrow et al., 2011). Social media addiction (Sun & Zhang, 2021) is another phenomenon much discussed among psychologists and sociologists linked to theories of hyper presentation and framing of self (Maddox, 2017), fear of missing out (FOMO, Oberst et al., 2017), and privacy paradox (from the data privacy perspective, see e.g., Waldman, 2020). The problem is that all of these current conditions are connected to present concerns linked to the recent infiltration of AI throughout our global society. Google feeds AI and is AI. Human memory loss is Google's gain. Thanks to AI algorithms, social media gives us what we want—and a lot of it. AI reminds us when we are missing out, and despite the warnings (General Data Protection Regulation, GDPR), we would still rather be in the domains, feeding our data, than being excluded from the rest of the post-post-industrial human kingdom. Moreover, as critical as we are of the new hit, Generative AI (GenAI— ChatGPT, Midjourney, Nightcafe, DALL E 2, Stable Diffusion, etc.), for all the above-mentioned reasons, many (if not most of us) love to use it for its output and convenience.

More problems are rearing their ugly heads, such as digital sweatshops in South-East Asia and Sub-Saharan Africa (again old wine in a new bottle in the Global South) to power OpenAI's 'ethical AI' data supplier Sama (Jason, 2023). Workers labouring on a menial wage with appalling conditions, scan labelled data of all kinds of dark content (child pornography, torture, incest, etc.) to train the language models to detect and filter before reaching the users of the Global North (see also Nyabola, 2023). All of these dilemmas stem from and revolve around profoundly human afflictions. As Sam Altman (OpenAI's founder) states, 'AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies' (cited in Jason, 2023), seeing that money wins over the logic of life. The question and trajectory of AI in itself—the re-creation or mimicry of human thought to replace and surpass humans in a variety of tasks—is characteristic of the human condition in which through pride, demonstration of prowess, control and dominance, we are willing to outsource ourselves. This particularly holds for an era in which

not only our technologies become ever more autonomous while ironically being connected, but our lives become ever more disconnected from one another—while ironically *being connected*.

Humaness and Ethics

The term 'humane' conjures many ideas of benevolence and compassion between human beings as well as from human beings towards other beings. Ultimately, humane is applied to describe actions and decisions that exhibit empathy and promote wellbeing towards others. It can be considered a proponent of *ethics* and morality in terms of behaving according to conceptualisations of right and wrong. It can also be understood as the act and agenda of putting others' best interests and values first—i.e., not prioritising profit over human and environmental wellbeing. Paul Goodman (1969) in his essay, 'Can Technology Be Humane?' explains human commitment to technology. Reflecting on the historical Hiroshima bombing (1945), Goodman emphasises the commitment human societies give towards technology as if it were a religion. Technology as truth—a product of objective science enabling objective and accurate observations and results, and a means of semantic distinction (i.e., social connection to technology and products seen in brand identification, etc.)—is a silent chant that has characterised the modern era (see Danaher & Saetra, 2022; Drew, 2016).

Goodwin (1969) argued that people need 'meaning systems' in which to believe, yet already during the 1960s technology in its entirety, or technological determinism as a means of *definition*, had already overtaken the human and the spiritual. This can be seen as one of the current 'sore points' of current ethical AI dilemmas, for we know through, e.g., GenAI that the phenomena we relate to via the technology, and often the technology itself, is not what we think it is (i.e., deepfakes, hallucinations and blackboxes). Thus, in his foresight, Goodman argued for a cybernetic approach towards techno-societal development in which technology would be studied and developed with humans and their values in mind.

Ethics is a broad philosophical field that deliberates questions of good and evil, right and wrong (see e.g., Saariluoma & Rousi, 2020). It relates to morality, moral behaviour and moral decision-making, that either: (a) is used descriptively to indicate particular codes of conduct that are promoted by societies or communities, as well as individually; or (b) normativity, to indicate codes of conduct that are promoted universally

on the basis of rationality (Gert & Gert, 2020). Ethical considerations refer to sets of principles and values that pertain to matters of good or bad in relation to humans and their actions (Reeve, 2014).

Ethical considerations when deliberating social-technological concepts is one step towards transforming the human to the *humane* (compassion and benevolence). It is the humane, and what we term 'humane-centred design' that we focus on in this edited book. The book features the thoughts and studies from prominent international scholars, who from their own positioning, case and context, approach the *humane* and its potentialities in autonomous technology. The featured chapters dance across disciplinary boundaries to provide deeper, systemic insight into the factors, states, challenges and opportunities of AI and related autonomous technology in a variety of domains.

As editors, we have striven to select authors and studies that encompass the complex ramifications that technology of this nature (information-intensive, so-called self-learning technology) have on human life and experience, and how these technologies in turn, reflect the deeper darker workings of human cognition and values. Contributors represent a broad range of disciplines: consumer studies, marketing, design (service, user experience and interaction design, industrial design, illustration and game design), art (conceptual, drawing, performance, print-making, community and participatory), software engineering, sociology, cognitive science and communication studies. This broad range of disciplinary thinking is condensed into three distinct sections of the book (see Fig. 1.1).

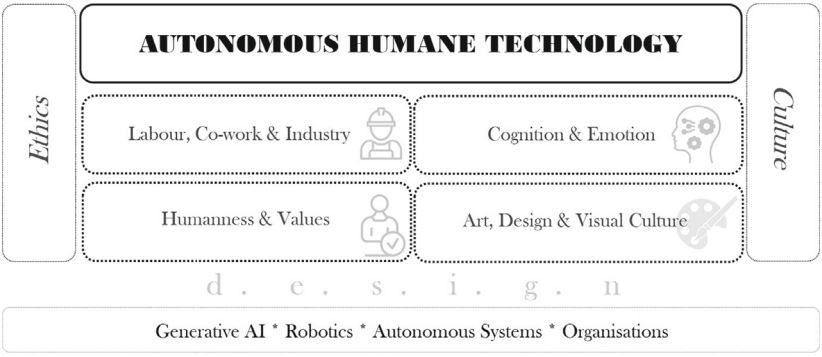

Fig. 1.1 Structure of the book

Structure of the Book

The book is divided into four themes (see Fig. 1.1): (1) Labour, co-work and industry; (2) Cognition and emotion; (3) Humanness and values; and (4) Art, design and visual culture (see Fig. 1.1). The themes capture the diverse areas of scholarship represented by the book, and highlight the ways in which AI-related phenomena are streaming through all sectors and disciplines of society. The AI revolution at workplaces has started, and we need to rethink how work with autonomous systems should be organised. Likewise, the nature of knowledge work and indeed understanding of creativity (in work, art and design) are being twisted by what we witness in the GenAI era.

Theme 1 focuses on various aspects of this workplace revolution that ever increasingly sees humans and AI as co-workers. In Chapter 2, Martina Čaić, Robert Ciuchita and Giuseppe Chiorazzo take a look at the dynamics and complexities of expert co-work between humans and chatbots. In their chapter, 'Help me help you: The dilemma of collaboration between expert employees and chatbots' Čaić and colleagues describe an illustrative case study in which semi-structured interviews were used to observe changes to human roles in B2B knowledge work. The chapter delves into the heart of current developments in the human-AI field, observing not simply an appropriation of traditionally considered *human qualities* in task distribution, but a more profound pondering of the evolution of human engagement in work, afforded by the AI systems. Yet, other factors arise that challenge us to critically evaluate the current and unfolding states of human-AI co-work that implicate our identities as human beings, including self-worth.

In contrast to passive monitoring of automation, in Chapter 3, Virpi Roto suggests that humane autonomous technology has the potential to enrich the work for humans. Roto's chapter, 'Co-worker, butler or coach? Designing automation for work enrichment', forces us to consider whether or not we want our AI to be designed to behave like a colleague, a butler serving us, or a coach that helps us grow as a people. The chapter draws on theories such as Self-Determination Theory (Deci & Ryan, 1985) as well as Eudaimonic Well-being (Waterman et al., 2010) to observe how AI can take workers to deeper stages of work enrichment over time. From another human-centred co-work perspective, in Chapter 4, Sarah Fox and Samantha Shorey observe how the implementation of AI affects the tasks and routines of janitorial staff and recycling

sorters. Their findings indicate a shift in tasks that transform practical workers into machine supervisors and technology trouble-shooters. On the basis of these empirical findings, they propose recommendations on how AI should be installed, evaluated and maintained in co-work environments.

Theme 2 of the book focuses on 'Cognition and Emotions'. The represented chapters apply understandings of cognition and emotions from the perspectives of: embodied technology design experience; human embodied expert cognition and the Most Advanced Yet Acceptable theory; and dual process theory within the context of technology-enabled services. Hanna-Kaisa Alanen and Rebekah Rousi's Chapter 5, 'Moving spaces—the affective embodied self in tram design and the autonomous imaginary' looks at current experiences of the emerging present—a newly established tramway in the city of Tampere, and how this is intertwined with the imagined future of autonomous transport. A symbolic interactionist (Blumer, 1980) approach is taken towards understanding how experience is formed across the dimensions of the material and immaterial in relation to self, identification of self and distinction (Bourdieu, 1979, 1996). Here, experience is formed in relation to the design and *others* as well as within the design. This notion of being 'within design' is taken up in Chapter 6 by Rousi who conceptually examines the Most Advanced Yet Acceptable (MAYA) theory (Loewy, 1951) in relation to disruptive autonomous maritime technology. Rousi focuses on the role of expert mental models and embodied cognition in ascertaining '*Whose MAYA*' is at play in popular perceptions of autonomous maritime vehicles and disruptive design developments.

Rolf Findrud and Bård Tronvoll examine human cognition in the unfolding era of human–machine co-work. By applying dual process theory to explore cognitive processes in the moments of, and in relation to, co-work with intelligent machines, Findrud and Tronvoll formulate a framework of human–machine symbiosis that incorporates necessary adaptation between human and machine 'thinking'. There is a relationship between Chapter 7 and Chapter 6 (Rousi) that see cognition, embodied affective processes, and the inner workings of human–machine synthesis as an integrated whole. In this integrated whole there is no superficial separation between humans and technology—they are intrinsically connected to one another, and in order to forge an effective and humane future, need to be developed in a way that compliments the various qualities presented by the organic and artificial actors.

This relationship is teased out further in the theme focusing on 'Humanness and values'. In Chapter 8, "Smart home technologies: convenience and control," Nils Ehrenberg elaborates on the nature of smart homes from the perspective of power. The chapter presents empirical findings that highlight the dynamics not simply between humans and their smart environments, but also in relation to circumstantial factors such as home renting versus ownership. This aspect is widely overlooked in the fields of human–computer interaction and computer science (Internet of Things), yet have severe ethical implications from a range of perspectives, none-the-least privacy. Ehrenberg applies a Foucaultian lens in what he describes as 'Panopticons of convenience' in which individual homeowners trade-off privacy for the perceived benefits of services, albeit 'autonomous' systems in the home (robot vacuum cleaners, mops, Alexa, etc.). The chapter challenges who the *actual users* are of the smart home, whether they are in the house or not. Users are subsequently divided into various types that once more instils a power structure that rests both between humans and technology, as well as between humans according to their role in the technology utilisation process.

In Chapter 9, 'Social robot design and the aesthetics of imperfection', Rousi and Paul Haimes explore the aesthetic possibilities of disrupting perfection. Where one technological trajectory has been that of human enhancement—the quest towards *human perfection*—Rousi and Haimes draw on the Japanese *Wabi-Sabi* and Christian philosophy to understand the *humaness* of the 'glitch'. They reach beyond the pragmatic to understand a social and psychological role of robotics that can only be satisfied through design that goes beyond skin deep. We all know how it feels to be lied to. On the topic of imperfection and then, potential deviance, Lars Witell and Hannah Snyder's Chapter 10, 'Dishonesty through AI: can robots engage in lying behavior?' draws on recent developments such as those seen in generative AI, in which the systems can be perceived as 'lying' to us. They refer to the inaccuracies and hallucinatory effects established when humans cannot rely on the algorithms they have created. More pertinently, the chapter focuses on the business and organisational context in which the ripple effects of content falsification (mis or dis-information) can be catastrophic. This particularly holds from the customer point of view in which trust is the maker or breaker of any individual-organisational relationship. Here, emphasis is placed on lying behavior in service encounters.

Four chapters on the theme Art, Design and Visual Culture constitute the final section of the book. First out, in Chapter 11 'Grasping AI entanglements—digital feminism and generative AI,' Tomi Slotte Dufva makes way for discussions on art and creativity. Slotte Dufva applies feminist phenomenology to analyse machine learning models and how they manifest predominant ideologies in AI-generated imagery. Moving beyond mere discussions on bias in the algorithms, Slotte Dufva challenges the reader to consider the broader implications of engaging with the technology from ecological, political and ideological standpoints. The author observes changes of discourse and roles in creative practices. In particular, he emphasises the dynamics of positioning humans in creative processes—from mark maker, image taker (photographer), image creator to 'prompt engineer'. Assemblage theory and entanglement are applied to understand the interrelationships and fragile dependencies between humans and technology in image production and subsequent interpretation. Prompt engineering in itself is textually and linguistically intensive, rendering a major level of abstraction from the output produced. Thus, the fabrics in between intention, input and output comprise intricate components of a larger assemblage that is ever more exposed to mathematical calculation and probabilities.

Aaron Knochel, in Chapter 12, 'Scribbles, spirographs, and AI, oh my!' advances this discussion by re-framing mark-making practices through graphicality, socio-materiality, meaning and participatory cultural engagement. Once more, Assemblage Theory (Delanda, 2006) is used to understand the social ontology of creative artistic practice, and practices of graphicality, yet this time in light of the *more-than-human*. It is a re-defining of autonomous technology through something that goes beyond ourselves. The chapter presents an alternative view on the potentials of creative production and human–machine symbiosis that is not rigidly tied to current developments in AI. However, with this said, the chapter aids in the pondering of what AI means to art practice, creative expression and human–human (multi-people collaboration) as well as human-non-human creation—as a tool and as an actor. 'Mr Fusion and Johnny 5? Visual rhetoric of AI design' (Chapter 13) by Stuart Medley and Jo Jung, takes us on a ride through time, popular culture and Modernism to discover 'the language of AI' in design. The chapter observes the human dimensions of trust and control in AI through visual aesthetics and rhetorical use of Modernist style. Modernism is examined as a stylistic metaphor used strategically in communication design related to and created from

AI and its generative AI counterparts. To illustrate this argument Medley and Jung analyse the design of several digital products, which harness the Modernist design philosophy of clarity, simplicity and understatement. Modernist design philosophy connects with historical movements attempting to unify through design, as a sign of progress and expression of 'honesty'—*form follows function* (Crouch, 1998). Historically, some notable moments of Modernist design history such as those seen in Scandinavian and Nordic design—post-world war two—correspond with critical advancements in AI development and conceptualisation. This particularly stands for major steps in the scholarship of cognitive science. As we see however, Modernism as a style or philosophy, can be seen as a smokescreen for complexity, manipulation and hypocrisy.

Finally, Chapter 14, 'The cultivated practices of text-to-image generation' by Jonas Oppenlaender, sees critical issues raised relating to the recent rise in popularity of AI image generation software. Oppenlaender closely scrutinises the co-creative ecosystem of humans and text-to-image generators by highlighting controversies such as intellectual property issues and the problems of rapid utilisation and spreading of synthetic data. The chapter is a gripping end to this edited book that has covered the opportunities and the potential ethical pitfalls of AI and autonomous systems becoming a normal part of our everyday lives. The aim has been to assemble the expertise and insight of scholars across the disciplines from service design, human–computer interaction, computer science, cognitive science, communication studies, visual art, and design in order to re-frame the significance of AI in our emerging culture. We have used these contributions as an opportunity to grasp the complexity and steer future AI design development towards one that embraces human values, and overall humaness—benevolence, empathy and care. Our intention is to move forward in a world where AI is a co-inhabitant. Yet, at the same time, our goal is to ensure that the co-inhabitant is designed and developed in a way that places people first.

References

Blumer, H. (1980). Mead and Blumer: *The convergent methodological perspectives of social behaviorism and symbolic interactionism* (pp. 409–419). *American Sociological Review*. https://www.jstor.org/stable/2095174

Bourdieu, P. (1979/1996). *Distinction: A social critique of the judgement of taste* (8th printing). Harvard University.

Chen, H., & Magramo, K. (2024). *Finance worker pays out $25 million after video call with deepfake 'Chief Financial Officer'*. https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Crouch, C. (1998). *Modernism in art, design and architecture*. Bloomsbury Publishing.

Danaher, J., & Saetra, H. S. (2022). Technology and moral change: The transformation of truth and trust. *Ethics and Information Technology, 24*(3), 35.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.

Delanda, M. (2006). *A new philosophy of society: Assemblage theory and social complexity*. Bloomsburg.

Drew, R. (2016). Technological determinism. In G. Burns (Ed.), *A companion to popular culture* (pp. 165–183). Wiley Blackwell.

Elish, M. C. (2019). *Moral crumple zones: Cautionary tales in human-robot interaction* (pre-print). Engaging Science, Technology, and Society (pre-print).

Gert, B. & Gert, J. (2020). The Definition of Morality. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2020 Edition)*. https://plato.stanford.edu/archives/fall2020/entries/morality-definition/

Goldman, J. (2023). *2023 saw a number of AI scandals, demonstrating the need for clearer guidelines for brands and publishers*. https://www.emarketer.com/content/2023-saw-number-of-ai-scandals-demonstrating-need-clearer-guidelines-brands-publishers

Goodman, P. (1969). Can technology be humane? *The New York Review of Books, 13*(20), 199–215.

Heath, A. (2024). *Google CEO says Gemini AI diversity errors are 'completely unacceptable'*. https://www.theverge.com/2024/2/28/24085445/google-ceo-gemini-ai-diversity-scandal-employee-memo

Jason, P. R. (2023). *Digital sweatshops and dirty AI* . https://jasoninstitute.com/digital-sweatshops-and-dirty-ai/

Loewy, R. (1951). *Never leave well enough alone*. John Hopkins University Press.

Maddox, J. (2017). "Guns don't kill people… Selfies do": Rethinking narcissism as exhibitionism in selfie-related deaths. *Critical Studies in Media Communication, 34*(3), 193–205.

Nyabola, N. (2023). *ChatGPT and the sweatshops powering the digital age* . https://www.aljazeera.com/opinions/2023/1/23/sweatshops-are-making-our-digital-age-work

Oberst, U., Wegmann, E., Stodt, B., Brand, M., & Chamarro, A. (2017). Negative consequences from heavy social networking in adolescents: The mediating role of fear of missing out. *Journal of Adolescence, 55*, 51–60.

Reeve, C. D. C. (2014). *Nicomachean ethics*. Hackett Publishing Co.

Rousay, V. (2023). *Sexual deepfakes and image-based sexual abuse: Victim-survivor experiences and embodied harms* (Doctoral dissertation). Harvard University.

Saariluoma, P., & Rousi, R. (2020). Emotions and technoethics. *Emotions in Technology Design: From Experience to Ethics*, 167–189.

Satter, R. (2023). *FBI says artificial intelligence being used for 'sextortion' and harassment* . https://www.reuters.com/world/us/fbi-says-artificial-intelligence-being-used-sextortion-harassment-2023-06-07/

Singer, N. (2024). *Teen girls confront an epidemic of deepfake nudes in schools* . https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778.

Sun, Y., & Zhang, Y. (2021). A review of theories and models applied in studies of social media addiction and implications for future research. *Addictive Behaviors, 114*, 106699.

Waldman, A. E. (2020). Cognitive biases, dark patterns, and the 'privacy paradox.' *Current Opinion in Psychology, 31*, 105–109.

Waterman, A. S., Schwartz, S. J., Zamboanga, B. L., Ravert, R. D., Williams, M. K., Bede Agocha, V., Yeong Kim S., & Brent Donnellan, M. (2010). The questionnaire for eudaimonic well-being: Psychometric properties, demographic comparisons, and evidence of validity. *The Journal of Positive Psychology*, *5*(1), 41–61.