AI & SOCIETY
https://doi.org/10.1007/s00146-024-01881-8

OPEN FORUM

Narrativity and responsible and transparent ai practices

Paul Hayes[1] · **Noel Fitzpatrick**[2]

Received: 19 October 2023 / Accepted: 17 January 2024
© The Author(s) 2024

Abstract
This paper builds upon recent work in narrative theory and the philosophy of technology by examining the place of transparency and responsibility in discussions of AI, and what some of the implications of this might be for thinking ethically about AI and especially AI practices, that is, the structured social activities implicating and defining what AI is. In this paper, we aim to show how pursuing a narrative understanding of technology and AI can support knowledge of process and practice through transparency, as well help summon us to responsibility through visions of possibility and of actual harms arising from AI practices. We provide reflections on the relations between narrative, transparency and responsibility, building an argument that narratives (about AI, practices, and those persons implicated in its design, implementation, and deployment) support the kind of knowing and understanding that is the aim of transparency, and, moreover, that such knowledge supports responsibility in informing agents and activating responsibility through creating knowledge about something that can and should be responded to. Furthermore, we argue for considering an expansion of the kinds of practices that we might legitimately consider 'AI practices' given the diverse set of (often materially embedded) activities that sustain and are sustained by AI that link directly to its ethical acceptability and which are rendered transparent in the narrative mode. Finally, we argue for an expansion of narratives and narrative sources to be considered in questions of AI, understanding that transparency is multi-faceted and found in stories from diverse sources and people.

Keywords Ricoeur · Narrative · Virtues · AI · Ethics · Responsibility · Technical practice · Transparency

1 Introduction

Artificial Intelligence (AI) has become subject to increasing scrutiny and conversation in recent years, and probably even moreso in recent months with tools (such as ChatGPT) regularly attracting mainstream media reporting and even notably capturing the imaginations as well as reasonable and concrete fears of those in the arts and entertainment industries (Donnelly et al. 2023; Nightingale 2023).[1] There has been a great deal of work undertaken globally in trying to manage the ethical development of AI tools as evidenced by the wide proliferation of AI ethics principles and guidelines, as well as regional efforts to regulate it and protect individuals and society from potential harms (see the European Union's upcoming AI regulation) (Hickok 2021).[2] AI holds a significant place in the public consciousness, and whilst it sometimes operates invisibly across many contexts and may not be noticed or thought about by those who interact with it (recommender systems for example), it is also something that is explicitly embraced and integrated into peoples' personal and professional lives and practices, and therefore shapes (and is shaped by) those lives and practices in the process of being put to use towards different ends. Because of the profound influence of instances of AI on how we live, of how we organise and get through our day or do our jobs, and indeed decisions that are made about us by public services and corporations alike, AI is something that is ethically charged with significant value (and virtue)

✉ Paul Hayes
paul.hayes@tudublin.ie

Noel Fitzpatrick
Noel.fitzpatrick@tudublin.ie

[1] TU Dublin, ECT Lab+, SFI ADAPT Research Centre, Dublin, Ireland

[2] TU Dublin, GradCAM, ECT Lab+, SFI ADAPT Research Centre, Dublin, Ireland

[1] See South Park Season 26, Episode 4, 'Deep Learning' as one example of this. ChatGPT is also credited as a co-writer of this episode.

[2] See https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Published online: 25 February 2024

implications, and implications for meaning-making and understanding ourselves and our society as we continue to move forward with new and evolving technological affordances.

Recently, scholarship has turned the lenses of (digital) hermeneutics towards technology generally and AI in particular in an effort to understand how we interpret ourselves and our technologies, and how we come to be together, and the ethical implications of this (Reijers and Coeckelbergh 2020; Fitzpatrick 2021; Kudina 2021). This hermeneutic turn provides a promising route for exploring what it means to be human in the digital age, with technology and AI, and how we can understand ourselves as responsible and ethical agents living together with technological affordances with which we co-shape the world and ourselves. This paper will build upon recent works in narrative theory and technology by examining the place of *transparency* and *responsibility* in discussions of AI, and what some of the implications of this might be for thinking ethically about AI and especially *AI practices*, that is, the structured social activities implicating and defining what AI is. This paper builds in particular on the works of Reijers and Coeckelbergh (Reijers and Coeckelbergh 2020; Coeckelbergh 2021b) in using some of their developments of narrative and process theories of AI as a primary framework for discussion of concepts of transparency and responsibility (both moral and narrative, as we shall see).

In what follows, we hope to show how pursuing a narrative understanding of technology and AI can support knowledge of process and practice as a source of transparency, as well as help summon us to responsibility through visions of possibility and of actual harm. The paper will proceed as follows. Firstly (Sect. 2), we will provide an overview of the different senses in which artificial intelligence can be discussed, building up to overviewing an encompassing view of AI as a process and narrative. Following this (Sect. 3), we will begin to unpack more what our overall conceptual and theoretical framework consists of (narrative and virtue theory) and detail more specifically what AI as narrative and process is. The subsequent task (Sect. 4) will be to provide an overview of narrative and moral responsibility and briefly analyse their relations. In Sect. 5, transparency (and AI transparency) will be overviewed, including its ethical relevance (moral transparency). With this work done, Sect. 6 will provide reflections on some of the relations between narrative, transparency and responsibility, building an argument that narratives (about AI, practices, and those persons implicated in its design, implementation, and deployment) support the kind of knowing and understanding that is the aim of transparency, and, moreover, that such knowledge supports responsibility in informing agents and activating responsibility through creating knowledge about something that can and should be responded to. Sect. 7 concludes the paper with further reflections on AI practices and narratives, emphasising the need to broaden our understanding of what an AI practice is, and whose narratives matter and why.

2 Making sense of artificial intelligence

There are multifarious ways of conceptualising artificial intelligence (AI), whether that is as a field of research, an object, a practice (or series of practices), or as a process (which also implicates practices) and/or narrative. In this section, we will capture some of these different conceptualisations of AI and the varying elements to which it refers, from specific data science and computational aspects to a broader socio-technical conceptualisation. We settle on a process/narrative conceptualisation as proposed by Coeckelbergh (2021b), one which encompasses many of these pieces and places them into a more comprehensive whole, briefly introducing this account for further elaboration and exploration of its implications for transparency and responsibility in the sections that follow.

From a data science and computing perspective, John Kelleher (2019, 251), argues that we may refer to AI broadly as '[t]he field of research that is focused on developing computational systems that can perform tasks and activities normally considered to require human intelligence'. Others argue that we may go beyond this to acknowledge different proposed goals of this field of research (Bringsjord and Govindarajulu 2022 citing Russell and Norvig 2016). A distinction can be noted between AI research that aims at AI that can act, as in execute actions, and reasoning (or 'thinking'), however, these categories are not incommensurable (Bringsjord and Govindarajulu 2022).

At even greater levels of specificity in the computing context, discussion may move into methods and techniques of AI, or different areas of research subsumed by it, or the various different applications of AI systems. Fundamental to the development of AI systems are algorithms, which are essentially sets of instructions which '[…] transform some data, which describe a problem, to some form that corresponds to the problem's solution' (Louridas 2020, 20). The design and development of algorithms is an essential element of the field of machine learning (ML), which focuses on the design and evaluation of algorithms '[…] that enable computers to learn from experience' (Kelleher 2019, 253). In ML useful patterns are extracted from data sets, with an ML algorithm (trained on historical data) that takes this input and returns a model (a computer programme) '[…] that encodes the patterns the algorithm extracted (or learned) from the data' (Kelleher 2019, 253). The scope of research in ML methods can be quite complex, with subfields that entail research into neural networks (deep learning), which feature multiple layers of neurons or hidden units (where each neuron is itself an

information processing algorithm '[…] that takes a number of numeric values as input and maps these values to a high- or low-output activation' (Kelleher 2019, 254). Instantiations of AI utilising algorithmic systems are commonplace, used across a variety of domains and contexts and are used in everything from judicial processes, policing, welfare decisions, recommender systems, and internet searches to AI assistants such as Alexa or the already mentioned generative AI tool ChatGPT.

When discussing AI we can move from the high level of the field of research to the individual aspects of the design and deployment of algorithms and artificial intelligence to discuss the general area and its specific features within the disciplines or practices of computer and data science, however, it is also possible to speak of AI from yet a more total socio-technical or systems-based perspective. Kate Crawford (2022, 8) for example is careful to describe AI in a more encompassing or total light, as something (or more than merely the singular thing) deeply embedded across social, political, cultural, economic and even ecological strata, arguing that '[a]t a fundamental level, AI is technical and social practices, institutions and infrastructures, politics and culture' and '[…] artificial intelligence is both embodied and material, made from natural resources, fuel, human labor, infrastructures, logistics, histories, and classifications'. Others go a step further when conceptualising technology today, arguing that artificial systems (which we could argue include instances of AI and its whole accompanying practices and infrastructure) are '…planetary exorganisms—both organized and organizing—within which and through which we live both as individuals and as political communities' (Krzykawski and Lindberg 2021, 195).

Most pertinently here, acknowledging all of these ways of conceptualising AI to some degree, especially the more socio-technical and encompassing perspective, is Mark Coeckelbergh (2021b), who uses narrative and process theory to conceptualise AI. Following in traditions of phenomenology, post-phenomenology and hermeneutics (Ihde 1990; Heidegger 2013) and owing to philosophical hermeneutics of Ricoeur in particular (1990a, 1990b, 1990c, 1994), Coeckelbergh (2021b) resists viewing technical artefacts like AI as mere objects or as things-in-themselves, but rather as processes and narratives linked to human experience and becoming (as Ihde (1990, 69) argued, there are no things-in-themselves, only things in multiple contexts). AI is the stories we tell about it, our individual and collective experiences with it, as well as being both a process and outcome of processes, it is ambiguous but interpretable, subject to the ascription of different meanings but also, as co-narrator of human experience and events, a 'meaning-maker' (Coeckelbergh 2021b). More will be said on this in what follows, with the next task being to further outline interlinked concepts of narrative, process, and practice.

3 Narrative, technology, and artificial intelligence

A hermeneutic and narrative approach to interpreting digital technology including human- technology-world relations, self-understanding, and meaning-making with technology has become a burgeoning field of inquiry, opening up fruitful possibilities for philosophical and ethical analyses (Capurro 2010; Romele, Severo and Furia 2020; Coeckelbergh 2021b, 2021a; Fitzpatrick 2021; Kudina 2021; Reijers et al. 2021). Hermeneutics has been 'classically' defined, according to Romele, Severo and Furia (2020, 73) as, 'an art, technique, and technology for the (correct) interpretation of cultural productions, mostly texts. [And] In the twentieth century, hermeneutics became a philosophical movement dealing with interpretation and understanding as the main features of humans' "being-in-the-world"'. However, as we shall see, it has since been developed to consider 'being-in-the-world' with technology, and its uses have been variable and approaches not 'unitary' as such (Romele, Severo and Furia 2020, 74). Some such recent works in hermeneutics have notably followed traditions of phenomenology and post-phenomenology which focus on the interpretation of human experience, especially those related to perception and body (Ihde 1990, 21), following inquiries into technological mediation and investigating how people make sense of technologies and themselves in certain socio-cultural contexts, how they co-shape each other and how people relate to the world with technologies (Kudina 2021).

Don Ihde (1990) was among the first post-phenomenologists whose works represent an important milestone in the philosophy of technology. Ihde (1990) pioneered enduring scholarship on technological mediation and articulated clear frameworks for understanding technology-human-world relations (and the many configurations thereof), notably under the categories of embodiment relations, hermeneutic relations, background relations and alterity relations. For Idhe (1990) the multipurpose aspects of technology and technical objects are part of their potential hermeneutics horizon, his concept of multistability enables Idhe to conceptualise the potential multiple uses but also to have a limit to those interpretations which remains the stable core of the technology or technical object. Gilbert Simondon's (2017) concept of individuation of technical objects includes these potential interpretations which are embedded in the technical not simply in terms of usage, but a critique of the reduction of technical objects to their simple use, they are also embedded for Simondon in processes of co-individuation or collective processes or cultural context and appropriation. For Simondon (2017) this a process akin to metastabilisation in chemistry, one at a certain point in time metastability can occur.

Further scholarship and development of a post-phenomenology inclusive of a significant hermeneutic perspective has more recently been undertaken by Peter-Paul Verbeek whose work focuses on analyses of technological mediations and technologies as moral mediators. Verbeek's (2016) work has gone far to demonstrate the relevance of technological mediation to knowledge (how they help to shape our ways of understanding the world; ethics (e.g., the possibilities for ethical and unethical action and practice emerging from technological disclosure of the reality and the shaping of moral decisions); and the metaphysical (the shaping of experiences of transcendence for example). Verbeek's (2015) work has followed trends in emerging technologies and analyses new relations beyond those identified by Ihde, including cyborg, immersion, and augmentation relations, and investigates points of application of the influence of technology, and the different types of technological influences there are.

The post-phenomenological and hermeneutic approach emphasises the integral contribution of technological and digital artefacts to mediation in the world, and its relationship with the individuation of people (of becoming human in a technosphere as such) (Fitzpatrick and Kelleher 2018). Such analysis is ethically important, usually emphasising the non-neutrality of technological artefacts that can be appropriated in different ways (even beyond those intended by their designers), or inclined towards certain (ethical or unethical) uses, and their capacity for transforming and creating human (social and/or technical) practices (Ihde 1990; Capurro 2010; Kudina 2021), and providing a framework for examining such ethically charged human-technology-world relations, especially through consideration of contextual information (use-cases and socio-cultural backdrops) that inform (and are informed by) technical practices. Indeed, the non-neutrality of the technical object has been critiqued by demonstrating that technical objects are never neutral, they are placed in a pharmacological context, prominently by Bernard Stiegler—which is to say that the technology as pharmakon is both cure and poison (Stiegler 2013; Fitzpatrick and Kelleher 2018). If this is the case, it becomes all the more important to reflect on the pharmacology of the appropriation of technology (or even of adaptation to new and harmful technological circumstances), to find ways for technology to extend and support human action and human becoming towards the good, and not merely stunting and stupefying it (Stiegler 2013; Fitzpatrick and Kelleher 2018).

What we are interested in here, in particular, is work developed in recent years by Wessel Reijers and Mark Coeckelbergh (see Reijers and Gordijn 2019; Reijers and Coeckelbergh 2020; Coeckelbergh 2021a, 2021b), which builds on ethical and hermeneutic philosophies of Paul Ricoeur (1990a, 1990b, 1990c, 1994), Alasdair MacIntyre (2013), and follows in the footsteps of David Kaplan's (2006, 49) work on the relevance of Ricoeur to the philosophy of technology. Their work also builds significantly on that of Verbeek's, providing new (narrative) avenues for analysing technological mediations (or configurations), understanding that designing technology, like Verbeek (2015, 2016) suggests, is designing humanity itself, and by presenting a theoretical and methodological framework that opens useful avenues for new empirical approaches (through narrative investigations).

Kaplan (2006, 49), following Ricoeur for example, argued that the hermeneutics of texts applies to the world of action (or, perhaps to that notion of being-in-the-world), that actions like texts are readable and can be interpreted quite apart from the intentions of their authors. Then, much like action, Kaplan also argues that technology functions much like a text or action '(a) technology on this model is like a text: it is readable, with a meaning that is independent of the intentions of the original creators and users' (Kaplan 2006, 49). Notably, Riceour's narrative theory reveals accounts of things—the telling of the story reveals otherwise hidden actors and motivations, design choices and ethical consequences (Kaplan 2006, 50), and as such stories can reveal the construction of the self in relation to technology, that is, becoming human and understanding the self through technological practice and technological mediation (Kaplan 2006, 51; Fitzpatrick 2021).

Reijers and Coeckelbergh make significant contributions to what Kaplan began by developing a philosophy and ethics of technology based on the hermeneutic philosophy of Ricoeur and the narrative-based virtue theory of Alasdair MacIntyre, also following developments in post-phenomenology by scholars such as Ihde and Verbeek. Reijers and Coeckelbergh (2020 42) conceptualise narrative '[…] as a mediation of human experience and understanding of the social world through the process of emplotment, which is the organisation of heterogeneous elements in a meaningful synthesis'. The paradigmatic model of the narrative is a text, which they argue similarly to Kaplan, is not restricted only to text but also mediates, importantly, technical practices in the world of action (Reijers and Coeckelbergh 2020, 43). Their aim is to use narrative theory as one of '[…] technological mediation that conceptualises how our understanding and actions are mediated by external things (texts, but also other technological objects)', though they move beyond Kaplan's initial work in this area and argue that their approach is not simply about understanding how we make sense of technologies and our lives in narrative ways, but that technologies configure human actions and life narratives (Reijers and Coeckelbergh 2020, 44, 55). Reijers and Coeckelbergh (2020, 44) argue that their approach to narrative theory respects the materiality of technology and technological mediation (something called for by Kaplan (2006) in his work) but also brings with it language (narrative),

temporality (in terms of emplotment and narrative ordering of events, and so forth), and the social (narratives configure social reality through configuring plots and refiguring social events).

Narratives capture (and organise) agents, patients (recipients of moral actions), things, and their interactions across time and space—they are rich in detail. Not only this, but they capture the goals and ideals of their protagonists or characters. Moreover, narratives are, with regards to emplotment, teleological, recalling that '[e]mplotment designates the organisation of events by which people represent action in a plot'—and the plot moves towards a conclusion (Reijers and Coeckelbergh 2020, 82). Emplotment is constituted of three phases of movement which shape 'human experience' and our 'understanding of temporal existence' (Reijers and Coeckelbergh 2020, 82). Each phase corresponds to phases of engagement with a 'text' (though extending beyond that into the social world, or world of action) and the transformations of understanding that may consequently occur. In the kind of narrative philosophy developed by Ricoeur (1990a, 1990b, 1990c) and applied towards the ethics of technology by Reijers and Coeckelbergh (2020, 83), it is phases of mimesis (prefigurative, configurative, and refigurative) that capture the process of engaging with a new text (or technology) where events are configured by texts (or technologies) and lead to transformations of prefigured understandings of the world.

Reijers and Coeckelbergh (2020) have undertaken a significant task in reformulating and further developing Ricoeur's hermeneutic philosophy and 'little ethics' for use in the domain of the philosophy and ethics of technology today.[3] One particularly significant contribution offered is their Ricoeurian framework of four key concepts to study technological configuration, namely: textuality, literacy, temporality, and distancing (Reijers and Coeckelbergh 2020, 91–106). Reijers and Coeckelbergh (2020, 104–105) explain that textuality 'shows us *where to look*', whether an object is more or less textual determines the kind of narrative we look at—one about something as prefigured and how it mediates practices (a bridge), or a high textuality object such as a car with which we are co-authors of a narrative, which features action-chains and constitutive rules (to drive) and standards of excellence (to drive safely and sustainably), and can link to life plans (such as being a taxi-driver as the authors suggest), in which case the narrative is nested 'with regard to the narrative of a person's professional life'. Following that, literacy shows *who to look at*, for whom is a technology accessible (who is literate in its use and who can be involved in its technical practices?), and as such encompasses a process of stakeholder identification (Reijers and Coeckelbergh 2020, 105).

Temporality and distancing both in different ways indicate the freedom of a human in technical practice or the extent to which they are author or co-author in technological configurations, and how they can link the practices to their ideals (of being virtuous in the practice), of how basic actions, ascending complexification, and those ideals connect (Reijers and Coeckelbergh 2020, 105–106). Where technological configurations are non-chronological (enabling non-linear choices by humans in terms of basic actions), and not abstracted (as in the functioning of an automatic car), for example, humans have the capacity to link the practice to their ideals, including by way of ascending complexification, and virtues or vice (Reijers and Coeckelbergh 2020, 106). With these concepts adapted from Ricoeur's work, we are provided with the conceptual tools to describe and examine technologies in technical practices in a structured way—they show us where to look and the extent of the influence of technical artefacts in different situations, how they may be invisible but significant, who uses and benefits from them (or suffers them), and so forth.

A form of this narrative technology approach can also be observed in Bruno Latour's work, who was also concerned with how persons and objects could issue and follow scripts (or programmes of action which can be inscribed in objects and are subject to new instructions) which positioned them as authors and characters and placed them in their degree of fidelity to the script either above or below it (Latour 2014; Future Learn n.d. 2024). This current iteration of narrative technology however brings with it the additional resources of Ricoeur's and Alasdair MacIntyre's theories that bolster it with a strong ethical and hermeneutic content and position narrative investigations within particular technical practices and within a clear framework that structures investigations around mimeses, time and space, and links them to human aspirations and dreams of the good life. Moreover, this approach may respond to Verbeek's (2016, 195) call for a hermeneutic that addresses mediations 'from within'(of 'perceptions and interpretations that can help to shape intentions and actions').

To return now more explicitly to AI, Coeckelbergh (2021b) as we already briefly mentioned, illustrates three ways that AI relates to time using process and narrative theory and therefore builds on the foundation of work undertaken with Reijers. Coeckelbergh (2021b) refers to AI in these relations on a more general scale, however, than at the level of *particular* technical practices. The purpose is not to understand particular use contexts of AI but to conceptualise it more broadly from a narrative/process theoretical framework as a starting point. AI as narrative or process is arguably the totality of stories and visions and movements, being and becoming with AI (and of AI and persons), if not

[3] See also (Fitzpatrick 2021).

all connected technical practices that constitute it (especially to the degree that those practices constitute processes). The three relations described by Coeckelbergh (2021b) are:

- *The time of AI*: this relation represents the stories we tell about AI at both macro (such as cultural imaginaries, stories about society) and micro levels (stories about one's life). It is a collection of narratives about human experience, which are ethically charged (e.g., stories of discrimination which are so prominent), and distant ideals and fears, which actively shape the direction in which AI moves.
- *AI in time*: this relation refers to use and development processes (e.g., data science process) occurring in time, in separate but merging accounts as scientific-objective time and lived time in the lifeworld (Coeckelbergh 2021b, 1628). Processes, such as data science, are broken down into steps (such as data collection, modelling and so forth), that are both measurable and quantifiable but also lived and experienced by humans (Coeckelbergh 2021b, 1629), who are shaped by the process. AI is both such processes and their outcome (Coeckelbergh 2021b, 1629).
- *AI-time*: AI-time refers to AI as an active narrator that shapes time (Coeckelbergh 2021b, 1629). AI 'acts as a time machine' for instance through classifications based on historical data, thus anchoring us to the past and influencing present and future, and through prediction influences future outcomes (or processes and practices) too—think of input data decisions (biased accounts of reality or simply inaccurate ones) motivating decisions in hiring or even in justice and security (Hayes, van de Poel, and Steen 2020; Coeckelbergh 2021b, 1629). Then AI may be a decision-maker, or itself a character or co-narrator in the story shaping narratives (and changing people's stories), processes, and outcomes (consider decisions on access to social welfare services) (Coeckelbergh 2021b, 1629). The AI at this point is both a structured process and a narrator of human lives (to the degree that it co-shapes them) (Coeckelbergh 2021b, 1630). AI and people are not fixed things, but emerge from processes they are in (AI emerges from data science processes) and roles they have in narratives which they are assigned (people emerge from processes as 'manipulated consumers' for example within 'marketing processes' and 'capitalist narratives') (Coeckelbergh 2021b, 1630).

Each step of the interconnected processes described represents a particular practice or technical practice (e.g., data science, marketing, or healthcare) in the becoming of AI (and humans). AI therefore is not only inseparable from its processes and narratives, but the practices it is generated from, which it influences and transforms. Reijers' and Coeckelbergh's (2020) novel approach to narrative and technology ethics provides a framework for investigating the particular technical practices with and of AI in different particular or linked contexts, and the degree to which these practices are technologically emplotted (or the degree to which AI as co-narrator influences it). A narrative investigation of technical practice can bring to bear the mechanics of practice and how technology changes it or transforms it (and ultimately our understanding of the practice, ourselves, technology, and the world). This is an ethically salient task, as practices figure into our life narratives in meaningful ways, and are designed to produce personally and socially meaningful or valuable results (internal goods). In the next section, we will outline further what we mean by practice, and why it is so ethically salient.

3.1 Practice (and its ethical substance)

The ethical relevance of practice is pointed to perhaps most clearly in the virtue ethics tradition, in which practice can be seen almost as a process of doing ethics in the world in the virtuous pursuit of the good life. Alasdair MacIntyre (2013, 173–174), following Aristotle, argues that humans have a *telos*—every human activity, practice, etc. aims at a particular good, it is inherent in our nature. Within this scheme fits the overarching *telos* of *eudaimonia* as a state of 'being well and doing well in being well' (MacIntyre 2013, 174). The virtues (dispositions to act and feel in certain ways) enable the achievement of *eudaimonia,* indeed living virtuously is the achievement of *eudaimonia* (Aristotle 2009; MacIntyre 2013). MacIntyre (2013, 218) is particularly concerned with virtues in practices, where practices are:

> […] any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended.

To explain this using a pertinent example, persons involved in the practice of data science may seek to reach and push standards of excellence (both in the very performance of the practice and its products) in modelling an algorithm that effectively detects trends or reveals insights into some phenomenon. A person involved in programming and coding will aim towards standards of excellence by programming a machine learning model successfully implemented by its code. Moreover, practices are not monolithic and are 'continually redefined, reshaped, and extended by […] practitioners trying to achieve its standards of excellence, ends,

and goals' (Chen 2015, 84), a process likely accelerated by the development of new technological tools transforming the essence of practice as time goes by. At the most basic level, practice consists of basic actions such as gestures (pressing a key on a keyboard), which extend across action chains within a framework defined by a practice's constitutive rules and standards of excellence (Ricoeur 1994). Constitutive rules give actions meaning within practices, it is what gives the action of digging a hole to plant a seed meaning in the context of farming, or pushing keys on a keyboard while programming an application in software development (Ricoeur 1994, 154). The definitive example is that of the meaning granted to pushing a pawn on a chess board granted by the constitutive rules of that game (Ricoeur 1994). In this context, if we can extend the metaphor of the chess game rules, we can note there is tension between how the technology has embedded behaviours or encoded behaviours which influence the way in which the narrative can be constructed. The individual narrative of ethical responsibility is not completely separated from the environment in which the narrative takes place, this is akin to the ancient Greek notion of *hexis* of Aristotle or more recently the concept of *dispositif* from Michel Foucault (1980), *dispositif* or environment influence the disposition of the person. In the context of AI we could think of how the use of AI such as recommender systems determines the choices and social network feeds of the individual, the individual ethical narrative is pre-determined by the *dispositif* being put in place. There is a need, therefore to include questions of explicability within the environment.

Ricoeur (1994,156) notes that the practitioner is connected with those that came before them, and through (learned) constitutive rules, they are rooted in traditions, which while they can be violated, must first be assumed (Ricoeur 1994, 156). Yet 'the practical field' does not develop linearly, but moves in a twofold movement of ascending complexification (from basic actions and practices) and descending specifications ('[…] the vague and mobile horizon of ideals and projects in light of which a human life apprehends itself in its oneness' (Ricoeur 1994, 158).[4] Linked with this, within the realm of praxis and associated with practices (as they relate to those practices we choose) are 'life plans', '[…] those vast [action configuring] practical units that make up professional life, family life, leisure time, and so forth' (Ricoeur 1994, 157). Such life plans take shape in a dynamic and ambiguous move towards distant ideals which must be specified by '[…]the weighing of advantages and disadvantages of the choice of a particular life plan on the level of practices' (Ricoeur 1994, 158). Then, linking back with MacIntyre's *narrative unity of life* (the basis for the aim of the good life), Ricoeur (1994, 158) argues that this is a result of summing up of practices in a global form also governed by life projects. It is a narrative theory that provides the impetus for the hierarchical formation of units of praxis, starting with practice, then life plans, and at the top the narrative unity of life, where this overall narrative structure provides the basis for giving ethical character to action, for providing the grounds for *self-esteem* (Ricoeur 1994, 158, 175).[5]

Now discussing action and practice once again in an ethical light, Ricoeur (1994, 176) argues that a crucial, ethically charged element of practices is standards of excellence, which characterise a particular practitioner as 'good'–these are evaluative rules of comparison '[…] applied to different accomplishments, in relation to ideals of perfection shared by a given community of practitioners and internalized by the masters and virtuosi of the practice considered'. Such standards emerge from the common culture of a practice, and lasting agreement on the criteria that define success and excellence in the practice (Ricoeur 1994, 176). Standards of excellence indicate the kind of goods internal to practices there are (and the *telos* of actions in practice) (Ricoeur 1994, 176), and link to the 'reflexive moment of self-esteem', as one can appraise their actions in light of standards of excellence (Ricoeur 1994, 177).

Practice is important to discussion here as it incorporates actions that utilise and are configured by technologies including AI applications, and AI applications emerge from practices (data science for example). We are especially interested in technical practices, those which are defined by their technological context (indeed, if any practice can be defined without some reference to technology (Ihde 1990)). Whereby a technical practice is embedded in the processes of becoming AI, where the narrative of AI shapes and is shaped by these practices of design, development, use, sustenance, and maintenance, such practices may be considered *AI practices.* Practices in their relationship with virtues and the pursuit of the good life, as well as their social nature, are inherently ethical (see Kudina 2021). Furthermore, Reijers

[4] For a detailed and insightful explanation of ascending complexification and descending specification, we should again refer the reader to Reijers and Coeckelbergh (2020, 73):

Ascending complexification denotes the movement from basic actions and practices towards ideals, for instance starting to play the piano at an early age, by pressing some keys, and moving towards the ideal of becoming a professional piano player. Descending specification denotes the movement from ideals towards practices and basic actions, as in the above-mentioned example of reading 1984, which enables the practitioner to start from ideals and move towards a change in practice and its related basic actions.

[5] Do note though that Ricoeur is critical of elements of MacIntyre's philosophy, for instance, that the narrative. unity of human life is not something which cannot exclude fiction nor can it be tidily coherent as it is an 'unstable mixture of fabulation and actual experience' (Ricoeur 1994, 162; Reijers and Coeckelbergh 2020, 74).

and Coeckelbergh (2020, 4–6), explain that through the narrative approach practices are made intelligible, the agency of 'actions-with-technology' is revealed, and in the narrative mode practices with technology take on significance, and it can be made clear how technologies promote or obstruct particular virtues in context. It is also important to understand practices because they have a pre-narrative quality, or *mimesis*, which means they prefigure narratives (Ricoeur 1994, 157).

Having provided an overview of several important concepts that will be useful to bear in mind going forward, we will now proceed to examine narrative and moral responsibility—these are both important for understanding meaning-making and the ethical content of narrative and its relation to practice and the obligations imposed upon us to responsibly create and live narratives that also sustain ethical practices.

4 Narrative and moral responsibility

Narratives describe ways of living and being, and are intimately connected with visions of the good life—the meaningful ways in which we attempt to direct our narratives as authors (especially through particular practices), which we may succeed or fail in, and, moreover, succeed or fail in with AI as co-narrators on that journey. Our stories then are value-laden (or perhaps our values are 'story-laden' as Donna Haraway put it (Ihde 1990, 215 quoting Haraway 1986, 78)), as are our stories with AI and technology. These narratives should be read, written and ultimately authored, not only to help us understand ourselves better but to understand how to author better stories—we have hermeneutic and moral responsibilities stemming from our ethical and epistemic obligations to live well with others, and understanding what exactly this entails, how to do it, and who we are. But what do we mean when we speak of responsibility, and how does it relate to narrative? There are multiple varieties of responsibility, and more recently even an account of narrative/hermeneutic responsibility has been elaborated by Coeckelbergh (2021a). These varieties of responsibility are not discrete, however, so it is useful for us to reflect on them a little bit to explore their normative implications going forward.

Ibo van de Poel (2011) provides an excellent overview and conceptualisation of responsibility in evaluating the relationship between forward-looking and backward-looking responsibility. In this account, backward-looking responsibility refers to notions of responsibility including blameworthiness and accountability. Here, backward-looking responsibility then refers to the situation where persons responsible for some X (seeing to some state of affairs) are put to account (they are answerable) or blame where the state of affairs X does not obtain—in being accountable the agent A provides an account of what happened and their role in its happening. Where an account is not satisfactory, agent A may be blameworthy—that is, where they had the capacity to influence the state of affairs, a causal role in its happening, there was wrongdoing involved, and they cannot be excused by factors including ignorance and lack of freedom (van de Poel 2011). Forward-looking responsibility refers to the moral obligation to 'see to it that' some X is the case (van de Poel 2011). In the case of relational responsibility, we can specify that Agent A is responsible for some X or not-X to Patient B (the recipient of some moral action, or person subject to some undesirable state of affairs X)(van de Poel 2011; Coeckelbergh 2020, 2021a). Such accounts of responsibility come under stress today with regards to extensive causal chains and the problem of many hands (van de Poel et al. 2015; Coeckelbergh 2021a), but are nevertheless essential points of entry to understanding taking responsibility and being responsible for something, regardless of how difficult it can be to attribute actions and wrongdoing to particular agents.

Returning to a virtue-based framework, responsibility may be considered a disposition which is a '[…] *readiness to respond to a plurality of normative demands*' (Williams 2008, 459)—therefore, responsibility as a virtue might be said to refer to the kind of exemplary Agent A that recognises some (plural) X and their role in its obtaining or not, as well as a willingness and ability to be responsible (and accountable) for it as well as recognise, negotiate, and dialogue around the competing claims by patients with a stake in X (or competing Xs) as mediated by given situations (Williams 2008).

Finally and crucially, there is of course narrative responsibility, or '[…] the responsibility to make sense, to interpret, and to narrate', which is distinguished from moral responsibility in not being owed to others as such but to ourselves as individuals and communities in particular contexts (Coeckelbergh 2021a). It is about sense-making of things that happened, or who we are or want to be as humans (Coeckelbergh 2021a). Hermeneutic or narrative responsibility can also be characterised in terms of being backward- or forward-looking. Where it is backward-looking, it may refer to accounts or narratives about things that have happened (such as newspaper articles about some accident), and in the forward-looking sense can refer to the imaginative work of meaning-making and crafting narratives of the future (which can shape the future) (Coeckelbergh 2021a).

So, in moral responsibility we have a responsibility to see to some X (and account for our role in its obtaining or not)—whereas in hermeneutic responsibility we have a general responsibility to ourselves and others to make sense of some X in a way that is achieved through narrative. The latter is not a morally neutral exercise, and one must note that narratives can emerge through processes of

accountability (Coeckelbergh 2021a). When one is put to account, and they provide an account (potentially beyond justifications or excuses), they are potentially explaining their role in something's happening, and hence perhaps why it happened, it contributes to sense-making, to *understanding*. Moreover, using our imaginative capacities to imagine and share futures worth having, or linking narratives from the past to the present and future, is an inherently ethical activity that represents our effort to link narratives to the good life. Hermeneutic responsibility would appear to be a responsibility to read and write (to write narratives and indeed to read them—to make sense of something), whilst moral responsibility is arguably the obligation of authorship itself, to actually seeing to the realisation of particular narratives. If hermeneutic responsibility concerns making sense of things through narratives per se, when we speak of moral responsibility (and recognising the narrative co-configurative potential of our technologies, especially AI), we speak of ensuring that we choose worthwhile narratives and manage as best we can how such narratives are co-authored with technology.

Returning again to practice, it might be said that it is our hermeneutic responsibility to make sense of practice (especially *technical* practice), to understand technological emplotment (using mimesis and the accompanying concepts of literacy, textuality, distancing and temporality), to make sense of how technologies configure practices or the narratives that configure and are configured by them, and whether practices with technology align with the plurality of visions of the good life. The three relations between time and AI again are relevant here—narratives of practice show *the time of AI*, can capture processes of *AI in time*, and can help us understand *AI time*. Moral responsibility lies proper in the next step, which is using our understanding to try to actively shape narratives and technical practices that can serve the plurality of visions of the good life.

5 AI and transparency

Having discussed narrative and responsibility, we can now begin unpacking how those concepts relate to transparency and subsequently explore some possible normative implications of this, particularly with regard to AI. Whilst conceptualisations of transparency can differ quite significantly depending on the context (whom is referring to transparency and what their particular interest or goal is), there are common properties to which we often refer when discussing transparency of something, for example, the availability of some information, its accessibility, understandability, and relevance to a given query (Turilli and Floridi 2009; Tu, Thomborson and Tempero 2011; Tu 2014; Hayes, van de Poel and Steen 2023). Arguably, transparency can be understood to obtain about something (X) where a query about it can be answered because a certain threshold of information supports knowledge and understanding around it (thus, transparency is also teleological). For our purposes here, a useful, synthetic account of transparency has been proposed by Hayes, van de Poel, and Steen (2023, 592), who argue:

> (1) that transparency is not an attitude but a state of affairs; (2) that we should understand transparency teleologically as always being of something (X), for some audience (A), and for some purpose (P); and (3) that transparency requires information that meets a number of attributes […] so that it can communicate knowledge that satisfies P, and ideally leads to understanding by A.

The above is a general conceptualisation of transparency intended to be applicable across a number of domains and disciplines, but one which was formulated ultimately with concern for addressing transparency of algorithms and AI. These authors also recognise that the significance of AI is not that it is a standalone object but something which exists within or as part of significant and complex socio-technical systems or assemblages (see also Ananny and Crawford 2018) and as such when speaking of transparency of AI, we are not (always) interested merely in opening a technical black box but we may have a variety of questions relating to different aspects of the assemblage of people, processes and organisations in which AI tools are embedded, and the socially, ethically, and legally significant interactions of the parts of these assemblages (Hayes, van de Poel and Steen 2023). Transparency can refer to AI in different contexts or at different levels of abstraction, from elements of design and development (code and algorithm), to institutional embedding and societal impact (who uses particular AI tools and how it impacts those it makes decisions about, for example) (Hayes, van de Poel and Steen 2023).

Whilst the referents of transparency may not be, in many cases, overtly ethical (or at least, our interest in knowing some X, even if ethically non-neutral, may not be motivated by the attainment of some ethically significant knowledge), transparency is normatively significant, and its referents can be explicitly ethical. When transparency concerns expressly ethical matters, it can be specified yet further and conceptualised as *moral transparency*, which is, as argued by Hayes, van de Poel and Steen (2023, 586):

> …[the] state of affairs that obtains when relevant and understandable information about some X is available and accessible to some target audience (A), so that this information is sufficient for A for the purpose (P) of providing an account about X's supportive or conflicting relationship with relevant values and goals.

In AI, such transparency is motivated by understanding and rendering apparent issues including the presence of bias within input data and algorithmic models, factors that lead to bias, and discriminatory consequences or practices that may follow (or are sustained through toxic feedback loops of data collection) implementation of AI tools in certain domains (hiring, policing, etc.) (Richardson, Schultz and Crawford 2019; Hayes, van de Poel and Steen 2023). Such issues cannot be understood or rendered apparent by limited investigations of technical objects, however. In order for true 'moral' transparency to be meaningfully achieved, we must trace through the processes that constitute AI, from its inception, design, development and its implementation. Each step, each practice, contains ethically meaningful information (on people involved and their motivations and goals, datasets used, intended uses of AI tools and so forth) and so 'moral' transparency of AI cannot be reducible to fixed points in time and processes in isolation.

Transparency holds an important relationship with different varieties of responsibility too, including hermeneutic responsibility, which will be revisited in the following section. Transparency, moral or otherwise, can be relational to the degree that it is owed from some responsible Agent A (perhaps an agency that uses an AI tool that makes decisions about access to social welfare services) to a querying Agent B or even Moral Patient A (an applicant who has been refused social welfare supports). Those agents who are morally responsible to see to some X also need to be properly informed about the capacities and limitations of their AI tools to use them properly (Coeckelbergh 2020; Hayes 2020), for example, police would want to be informed about the limits of AI tools that make predictions about criminality or crime rates before operationalising them and potentially bringing harm to historically marginalised communities. Indeed, they are also responsible for ensuring the tool they use is fit for purpose—something which may only be possible to determine when there is the visibility of the processes leading to the creation of the tool and actors involved (the datasets, organisations, their goals and motivations, even histories of controversy). Additionally, there is transparency's relationship with backward-looking responsibility, or more particularly, accountability. The provision of an account, by providing clarity on someone's role in an event, contributes to the transparency of a particular X. Moreover, this is not a unilateral relation, as the existence of information from plural or independent sources may also support accountability by identifying different agents that should probably be held to account (that may not have been immediately apparent) (Hayes, van de Poel and Steen 2023).

Having outlined an account of transparency and described some of its relations with responsibility, it is now insightful to return more clearly to the narrative. Discussing this with reference to transparency shows how the narrative itself contributes to and can help provide yet a richer understanding of transparency, and how this, in turn, can support or interact with our responsibilities.

6 Narrative, transparency, and responsibility

6.1 Narrative and transparency

When we conduct the kind of structured narrative investigations into technology as suggested by Reijers and Coeckelbergh (2020) we reveal accounts of relations between people and technological artefacts, how they figure into their lives and co-shape each other through appropriation (see also Kudina 2021), across multifarious use contexts (the network on technical practices in which an artefact is embedded). Through a structured narrative investigation we gain access and insight into who the stakeholders (characters) in relevant practices are (makers, users, and governors), the degree of visibility and concreteness of the technology in human-technology-world relations (for example through considering distancing or even literacy), the degree to which a technology narrates or co-narrates stories (textuality), and the degree to which it may order actions (temporality) (Reijers and Coeckelbergh 2020). Through *mimesis*, we can explore narratives for accounts of pre-figured understandings of the world, how events and persons are organised and configured by plots, and how prefigured understandings of the world are changed by new technical artefacts and practices (that is, technological mediation is revealed) (Reijers and Coeckelbergh 2020). The narratives that are open to investigation may be both first order (first-hand accounts of experiences of technical practices) or second order (revealing pre-figured time and for which sources can be cultural, technical, or academic) (Reijers and Coeckelbergh 2020, 163–165). Indeed, narratives suitable for investigation may even be the cultural products (such as fiction) that reveal people's hopes and fears about technology, or those which inspire and motivate particular people and communities to make different technological artefacts or work towards particular desired futures (Cave, Dihal and Dillon 2020; Sartori and Theodorou 2022).

Narratives are ethically relevant in their descriptive power, revealing the things that hold value to people, why they engage in certain practices and the internal goods produced by those practices, as well as the constitutive rules that define them (and the basic action chains forming those constitutive rules), the standards of excellence by which they may be evaluated, the virtues which sustain them, and the linkages of life plans to a narrative unity of life and whether specific technical practices support

the move towards the good life, or, the Ricoeurian ethical intention of *living well, with and for others, in just institutions* (Ricoeur 1994). Narratives contain details relevant to answering a variety of queries about something or a state of affairs due to their arrangement of characters, things, and events—they have explanatory power. In fact, it can be argued that various things can be made transparent in the narrative mode. By organising characters and events into a meaningful whole (towards some conclusion), a narrative can paint a picture (from at least a particular point of view) of human-technology-world relations (and particular outcomes of these) across time and space. Narrative captures the human experience that can be used to impart information to others—where the narrative is written and shared it fulfils some of the transparency's requirements, it makes information about something available, potentially accessible, and understandable, and relevant information can be derived from narratives to communicate some knowledge that satisfies a particular purpose for particular audiences. Narratives give meaning, make events intelligible, and it is this intelligibility that supports the aim of transparency.

A narrative may structure the events and persons within an ecosystem of actors (or characters) in algorithmic transparency—an individual (now functionally, A–*audience*) can learn from, for example, narratives surrounding the practice of banking and loan approvals which may be supported by AI decision-making tools, with the P of knowing why their loan application was rejected. A narrative about an 'AI end user' might reveal the level of involvement of human decision-makers, the rules they followed and the applicable standards of excellence involved in making a decision—the availability of a clear narrative may indicate the soundness of the process and whether there was any recourse for appeal. In fact, in the latter example, a fictional narrative of another kind may even support the kind of transparency that responds to the loan applicant's transparency requirements. In the arena of explainable AI, one method of interest supporting the explainability of an algorithmic decision is the *counterfactual or contrastive explanation*. The counterfactual explanation takes on something of a narrative form that focuses on an alternate account of the events leading to a decision made by the AI (or AI end-user) about our A, whereby a state of affairs is described in which the outcome is different (Miller 2017; Wachter et al. 2018), or some element of a state of affairs is changed such as, for example, individual group membership (which can be undertaken to assess possible instances of discrimination) (Loi, Nappo and Viganò 2023). So, in the last example, a fictional narrative can be presented to A that basically imagines a world where the loan was approved and illustrates what is different about it (perhaps it is something as simple as a higher credit rating on A's part). In these examples, A comes to some knowledge (needed for a particular purpose) because narratives are available that contain accessible, understandable information that is relevant to A's query. Transparency can obtain, in some cases at least, through the narrative mode.

The ethical salience of the narrative and its relation with 'moral' transparency becomes clear when we, once again, re-centre discussion on technical practices. Recalling that 'moral' transparency has been argued to be 'for the purpose (*P*) of providing an account about X's supportive or conflicting relationship with relevant values and goals' (Hayes, van de Poel and Steen 2023, 586), we can see the importance of the description of practices and narratives that may necessarily or contingently engage some technological artefact. By describing technical practices, which themselves may form units of processes (including recall, in the coming to be of AI), and capturing them in the narrative form, ethically relevant information reveals itself which can provide an account about whether the technical practice, or elements thereof, supports ethical values and goals. Narratives, for example, may show whether the design or use of technological artefacts supports or obstructs virtuous conduct (or runs contrary to deontological rules or moral norms), and therefore whether it harms or supports the production of goods internal to various practices (which we may plausibly define as values too where these goods are states of affairs for instance). Narratives allow us to understand standards of excellence, and whether they are adhered to in a technical practice, or even possible. Reijers and Coeckelbergh (2020) provide illuminating examples in their book *Narrative and Technology Ethics* in this regard, relating to automated airport border security crossings. To elaborate on their border security examples, consider a narrative about someone passing through an automated border security crossing who may be stopped for having incorrect documentation. We can note from the outset the role AI could have as co-narrator of this story for having a causal contribution to stopping the traveller, but what the story will also show is whether a border control officer can exercise their virtues accordingly by overriding the automated border stop, discussing the circumstances of the traveller's case with them and whether they can uphold relevant norms and values in the context of border security (security, but also the safety and dignity of the traveller who may be an asylum seeker). This narrative is rich in detail that supports inquiries of 'moral' transparency, the arrangement of plot and characters can illustrate whether the technical practice of border crossing with automated checkpoints supports the relevant internal goods and goals of the practice. Moreover, when relating such narratives which may arise from an intensive narrative investigation, an analysis of the textuality, literacy, temporality, and distancing draws our attention to the important facets of the narrative and practice—that is, the extent to which the inclusion of AI has transformed the practice or may yet

have the capacity to transform it; who uses and is affected by the use of the AI, how it changes or alters the ordering of events in border security; and the capacity and freedom of human decision-makers to intervene. By achieving this transparency, we can begin to discuss changes that need to be made to technical practices and their associated milieu in order for them to hold a properly ethical character and not violate extant norms or obstruct the exercise of virtues. Narrative, it might be said, not only shapes practices but in the narrative mode also renders them transparent, and practices being inherently ethical social activities engaged in for the benefit of communities and individuals are an intrinsically important object of 'moral' transparency, they are the seat of ethically relevant human-technology-world relations.

In the prior example, the narrative captures just one story against what might be a very storied history of the development and design of the artefact (the automated border crossing), as well as a tradition of the practice that it augments (border control). The narrative mode also captures processes across varying time horizons, demarcated by broad categories to the extent that narratives can capture the relevant variables of those milieu and the technical practices that constitute those processes. For example, processes within an 'AI Supplier' category may entail a variety of connected or interlinked practices including marketing, data collection, modelling, programming and so forth, with interlinking narratives framing or being framed by, to some degree, these processes. Moving on to the 'AI End User', the narrative mode can capture the processes surrounding the deployment of the tool and how it is embedded in the organisation, and what ends it is put towards. Here, narratives can capture processes of governance and use of the tool. Finally, looking at broader society, narratives can capture the experience of individuals and society in technical practice, as they adopt it and use it (say, integrating ChatGPT into daily practices), or are otherwise impacted by it (the border-crossing example). This is to say that interwoven narratives capture the interconnected practices which in turn constitute processes of technological (and human) becoming and the availability of these narratives to the 'reader' renders transparent ethically salient features of the stories across time and space, and interlinked milieu, that support the kind of ethical inquiry aimed at by 'moral' transparency. The greater access we have to these narratives and processes, the better we will be able to influence them.

It must be noted that narratives are not necessarily immediately available or accessible. Narratives, whilst they can be lived, must be discovered or recorded to be available and accessible to those with transparency requirements. Yet such narratives are necessary for a variety of reasons. This is why we have a responsibility to 'write', that is, to record, document and disseminate (appropriately—indeed, some people may justifiably not want some stories told) narratives in appropriate forms, and also to 'read' such narratives—whether for reasons of narrative responsibility or moral responsibility, but more on this shortly. Narrative investigations, the kinds of which that are proposed by Reijers and Coeckelbergh (2020) do provide methods for exploring available narratives, and helping to discover and potentially document them through stakeholder engagements—we can discover and unravel narratives through qualitative research methods, for example, interviews, focus groups and so forth, with makers, designers, and users (broadly construed) of technological artefacts.

Narratives surrounding the socio-technical context of AI can be enlightening, perhaps even crucially so, however, it should be stressed that they are no panacea for the intrinsic opacity of some algorithms and AI systems. For a narrative to be possible, something must be able to be made intelligible in language and, essentially, within a story of some kind. The massively complex, abstract, and mathematical nature of some AI systems may be so severe as to continue to resist efforts at explanation and illumination by the narrative approach. A wide variety of methods will always be invited to try to illuminate what seems almost non-illuminable to also begin trying to render obscure processes into the story. For this reason, the ongoing efforts of the xAI community and the various innovations this area produces will always be complementary, or in mutually supporting relations with, a narrative approach.

A final point to mention on the relationship between narrative and transparency is that to some extent narratives can cast some light on the future—they can combat what Shannon Vallor (2018) calls technosocial opacity (the great uncertainty of future technological adoption and evolution). Through narrative foresight (and indeed other culturally productive exercises more generally) we can imagine futures, or varying alternative futures, and in so doing make some effort towards challenging technosocial opacity (Milojević and Inayatullah 2015). In constructing stories around what is to come, we can attempt to reduce the uncertainty that comes with the future, especially if the purpose of this foresight is not only to see into a future but to challenge existing or outdated narratives and attempt to choose a future worth having based on more subversive narratives that challenge the status quo (Milojević and Inayatullah 2015). Such future visions render particular (unrealised) futures (and futures of technical practice) transparent, even if to a limited degree. Predicted futures and the futures we choose to try to author are by no means guaranteed, however when we discuss transparency we refer to a highly gradated concept (Hayes 2020)—to imagine different possible futures through narrative is still to cast some light on what could be, and even if that exercise is only mildly successful, the needle would have been moved on the gradient of opacity towards transparency.

6.2 Reflecting on the relationship between responsibility, narrative, and transparency

Having reviewed the concepts of narrative, transparency, and responsibility, we can now reflect on some of their relevant relations a little bit more, as well as the consequences of these relations with regard to AI. In the preceding, we have broadly discussed two kinds of narratives, overarching framing narratives (second-order) and lived and active narratives (first-order) (Reijers and Coeckelbergh 2020). Both kinds of narratives can serve different kinds of transparency, being rich in information necessary for answering different queries. Narratives, in their relationship with transparency, can both serve and activate our responsibilities too—the interactions between these concepts can be quite complex. In the following, some core relationships between the concepts will be sketched briefly. For the sake of simplicity, these relationships will be grouped together under three headings; *narratives and transparent practices*, *narratives and transparent futures and alternative states of affairs*, and *narratives and moral transparency*.

6.2.1 Narratives and transparent practices

This heading is so named as it refers to the capacity of narrative structures to explain practices (and technical practices) (Reijers and Coeckelbergh 2020)—narratives capture the characters and events surrounding practices, why people engage in them (they may relate how they factor into life plans), what ends they pursue, how they are conducted (their constitutive rules), and how they are done well (standards of excellence). By linking together the various narratives that connect practices, we can also begin to understand interconnected processes and the relations between people and artefacts (and their joint becoming) across time and space, including in a manner that can describe or explain what a particular instance of AI may be in the richness of its context and the socio-cultural backdrops it spans. Narratives show the roles that technical artefacts play in people's lives, and the extent to which in their appropriated uses they shape the narratives of people's lives. By looking at narratives, we can see how technologies change technical practices, and how they create new ones. Second-order narratives, as they appear in socio-cultural productions (if not economic and political ones) about technology also evidence the narrative frames that structure practices and shape them.

The narrative mode makes practices transparent, as we have suggested, laying them bare for inspection through first-order narratives that capture human experience and second-order narratives that indicate wider ideals and motivational grounds that structure the boundaries and *telos* of a practice. Thus through telling stories (or 'writing' them) and listening to those stories (or reading them) we begin a process of explanation, understanding and meaning-making—we use narratives to make meaning vis-à-vis technical practice itself transparent. We engage our narrative responsibilities to each other when we explore narratives about our practices with technology to understand those practices and what they mean to us. The transparency of practice bolstered by narrative is a meaning-making activity, and not yet strictly an ethical one without further reflection on implications for relevant values and goals (or the good life) of a technical practice, for example. Nevertheless, our narratives leave practices bare and open to further reflection across different lenses, open to different questions bounded by relevance and the *telos* of inquiry.

6.2.2 Narratives and transparent futures, and alternate states of affairs

As we have suggested, second-order narratives about something like AI can have a framing effect that spurs its evolution in particular directions, and thus potentially mould technical practices to some extent in the image they promise. Today, we might argue that hegemonic capitalist, libertarian or techno-optimist narratives are spurring forth technical practices in particular directions (and indeed, such narratives may be destructive and certainly antithetical to the pursuit of the good life (see Stiegler 1998, 2010, 2013)).

Such narratives make the future, or possible futures, of technical practice somewhat transparent and by simply 'writing' them (or imagining new ways of being-in-the-world with technology (Nascimento 2019, 25)) we may begin the process of 'authoring' them, that is, making them lived realities. The process of constructing second-order narratives is an exercise of hermeneutic responsibility, allowing us to make sense of the present with reference to the future (Milojević and Inayatullah 2015; Coeckelbergh 2021a), it is also our narrative responsibility to tell stories about futures worth having, and thus those that configure our practices today in pursuit of those futures, meeting them through the ethical intention of living well, with and for others, in just institutions (Ricoeur 1994). When we make possible futures transparent, and choiceworthy futures at that, ones which can be achieved with the support of technical practices, we activate a certain moral responsibility. When the story is written, the choiceworthy life envisioned, we become responsible for our movement towards that choiceworthy state of affairs and seeing to it that it comes to be. We become responsible authors by advancing and living out choiceworthy narratives. This becomes an interesting responsibility because it is both narrative and moral, to construct (or seek out) second-order narratives as alternatives to arguably poisonous hegemonic ones that champion extractivism and exploitation, and see to it that we change the present to secure better futures based

on those narratives we construct to make sense of ourselves as humans (Coeckelbergh 2021a).

The hermeneutic endeavour also points to the open-ended nature of the interpretative process, one which is constantly renewed towards a horizon which is ultimately the story of life that can only be completed after death. In terms of explicability and explainable AI this poses the question of the open-ended nature of possible interpretations. The horizon here is one of Don Idhe's mutlistability, the potential for possible further interpretations of the technical object, interpretations which are not simply limited to re-purposing of data but also the very constitutive of the black box. Hence the very limit of interpretation.

We must however temper our expectations with regards to technological foresight and transparent futures emerging from different narratives, as often when we discuss technologies we can get trapped in hype cycles of overestimating their capabilities in the short-term and underestimating them in the long term (see Amara's Law) (Fitzpatrick and Kelleher 2018), potentially leading us to a Collingridge dilemma in being blinded by misleading expectations based on misleading narratives. Indeed, the plurality of dubious narratives about AI, for example, would tend towards fogging a future in false expectations and misinformation (Schwartz 2018; Stern 2023). This emphasises the need for us to pay attention to plausible narratives about the future (Coeckelbergh 2021a), and choose the ones worth having, whilst never losing sight of the present.

Constructing and making real choiceworthy narratives and the transparency they can engender are forward-looking cases of responsibility. It is worth once again discussing the relation of the backward-looking element of responsibility in relation to narrative and transparency. Consider again the case of the counterfactual explanation, for example. The counterfactual explanation is essentially a narrative about an alternate state of affairs, that is to say, a description of events not as they are or were but as they could have been where some variable(s) had been different (the reality where a loan had been approved rather than denied). The construction of this alternate reality through the narrative mode helps us to make sense of the state of affairs as they are—in glimpsing the world of alternate possibilities we can understand better the world we live in and why it came (or comes) to be. In seeing the world where our loan was approved, we understand better the world where it was denied, and potentially the causal chain leading to that decision. This explanation to some extent satisfies our narrative responsibility by helping us to make sense of some situation, and indeed in generating information and supporting understanding about some problem over which there is a query (and a Moral Patient) this narrative contributes to some transparency. In providing an account, or justification for a given state of affairs such as they are (rather than any other way), the counterfactual explanation also satisfies accountability (again, backward-looking responsibility).

6.2.3 Narratives and moral transparency

Narratives can provide a strong foundation for 'moral' transparency, in both first-order and second-order forms. Narratives can indicate the things that people value, that they strive for and care about, including the *eudaemonic* goods that factor into people's life plans and which may constitute a *telos* of practice. Narratives arrange characters and events, illustrate their motivations and goals, explain the purpose behind their actions with regard to the goods they seek and are even capable of conveying or evoking emotions that are indicative of the value of particular things to those characters (Nussbaum 2003; Roeser 2017)(consider for example, to borrow again a border-crossing example, first hand-accounts of persons describing feeling demeaned at being stopped at an automated border crossing, something which might concretely render the value accorded to personal liberty by that person). Narratives make clear what are important to persons in and from different milieux, they render the things that people value transparent—they make them available, accessible, understandable to us and such that they can be placed in the context of relevant queries.

Returning once again to practices, narratives detail them and render them, as we have seen, transparent by showing us their practitioners, those affected immediately by them, their constitutive rules and standards of excellence and the relevant goods pursued (or the values at which they aim) by the practices, as well as the kinds of virtues a practitioner may need to succeed in their practice. By examining technical practices within the narrative mode we can try to understand whether particular tools are suited to the ends to which they are put and whether they support or obstruct the virtues of practitioners, compromise standards of excellence, or tend towards violation of given moral norms. Narratives can show points of failure in a process, within interconnected practices, that lead to harm, that have harmed somebody, and that require intervention or remediation. Narratives support 'moral' transparency by drawing our attention to what makes technical practices conducive to good or evil and in so doing call us to responsibility, the responsibility to intervene in a technical practice that does not result in some desirable state of affairs X—or we might say the responsibility to re-design the technical practice such that it allows virtuous practitioners to skilfully aim towards the good. Such is the responsibility of all those with a capacity to act, as we must act as communities in responding to harmful technical practices, but also especially practitioners in the process of appropriating new technological tools which may

fundamentally change how they see their practices through. It is a particular responsibility of the practitioner to know the limits and risks of their tools, of the harm they may do to others and whether their practice truly benefits from them. Such knowledge can arise from 'moral' transparency, itself supported by examining the narratives about the practices and technologies.

The foregoing should have successfully highlighted some key relations between narrative, transparency, and responsibility. So far, AI has been mentioned in connection with this only sparingly. How might any of this be relevant to AI and how we think about AI, or even do ethics and AI? Narratives capture information vital to understanding what AI is. They capture the motivations and ideals that drive it in particular directions, they can support understanding the interlinked practices that form the processes of which AI is (as well as the outcome of the same), they help to understand who undertakes those practices and why, and whom these practices effect or impact, and the degree to which they are changed or co-authored by technology. Narratives are a store of transparency for those with the right questions. The field of persons with these questions can be quite diverse—at almost every level of society persons have transparency needs about things that affect them and will have valid questions of varying depth (including algorithmic decision subjects, academics, businesses, policy-makers and regulators, etc.).

Transparency, as the light that shines with knowledge, can activate or support our responsibility—those who know can better act upon knowledge, and indeed must if they can where it is in service of the ethical intention. Understanding AI as the ambiguous and evolving narratives and processes that compose it, and the human experiences entailed by this and understanding the ethically charged nature of the interlinked practices this also entails, should encourage us to think about how we make sense of those practices, whose voices and experiences matter in narratives about AI and its supporting practices, and perhaps to think more expansively about what those practices are in a deeply materially and socially connected world. In the next section, we will reflect more on making sense of ethical AI practices and offer support for the arguments of expanding the range of practices and sources and types of narratives about AI.

7 Making sense of ethical AI practices with expanding narratives

7.1 Expanding practices

To use and shape AI responsibly we need for these processes to be transparent, as well as the practices that shape and are shaped by narratives and reflected in the narratives we tell. Knowledge obtained through narrative and transparency enables us to act responsibly and also call us to responsibility (by making something known that previously may not have been). Such processes (moreover, the practices that constitute these processes) must not be narrowly construed. Understanding AI as both process and narrative must not stop at consideration of business and data science processes, or the underlying narratives tied to these milieux. To truly understand AI, and to make sense of it ethically and how we may ethically and responsibly manage its development and use, we need to appreciate it on a broader level of materiality and embodiment. AI as a process, a result of process, and a narrative, is the entire dynamic socio-technical assemblage of people and things that contribute to its ambiguous state (in all their materiality and varying relations). This point, again, is made with great lucidity by Kate Crawford (2022) who in the *Atlas of AI* painstakingly details the many facets of AI and how it comes to be–beyond just code and application–by highlighting the hidden and often underpaid labour of human beings in training algorithms (see for example Amazon's Mechanical Turk), the ecological and human cost of the extractive practices that sustain it (consider lithium and 'conflict mineral' mining that sustain the material structures of AI), and the economic, social, and political narratives that direct its development and use, often where the power over narrative is concentrated in so few hands (also see D`ignazio and Klein 2020). The ethical problems associated with AI practices are not just those found squarely in data and discriminatory or dangerous use cases (policing or military for example), data science practices dominated by persons from specific, dominant, and privileged standpoints (see again D`ignazio and Klein 2020), but also an array of less obvious but completely intrinsic practices across different industries and political spheres. Whether or not we consider a particular instance of AI ethical or just therefore requires more than an evaluation of practices of design in the tech industry and deployment and implementation by end-users, but a deeper consideration of AI infrastructure, resource supply chains and the variety of practices found across those (such as mineral mining, transportation, and so forth). Ethical AI is not only a project of ethics and ethical or virtuous practice, but indeed a large scale political project.

Making sense of AI then requires a reading of broader narratives and an examination of wider practices than may be immediately intuitive. And this is necessary for 'moral' transparency of AI, for how can we evaluate the morality or ethics of a technology without a significant understanding of practices (as processes) that constitute it? It is questionable that there can be any design of virtuous technical practices (Reijers and Gordijn 2019; Reijers and Coeckelbergh 2020) that utilise AI tools if any of those practices connected to them through time and space rely on the exploitation of people and the environment. This marks a challenge to the

virtuous practice design/narrative technology ethics proposed by Reijers and Coeckelbergh—one which is primarily concerned with the design of virtuous practices with technology and requires the mapping of practices that use particular technological instances. It appears, especially when looking at some technologies like AI as processes and narratives, that one cannot simply map those practices implicating the use of the tool but must also trace back further to the practices that deliver the tools themselves, even if somewhat indirectly. AI tools not only change and shape new technical practices in their appropriation, but their existence and evolution also creates a backward pressure on the other practices (extractive and so forth) that sustain them—the coming to be of AI tools re-patterns the practices and infrastructure through placing new demands on them. The technical practices that more explicitly implicate AI tools cannot be designed virtuously unless we consider how to decouple their coming to be from harmful, unethical practices that come before them but sometimes necessarily sustain them. AI then is not simply co-narrator of our lives through the technical practices that it supports, but its mere existence to some degree makes it something of 'ghost'-narrator in the potential configuration of the extractive, logistic, infrastructural practices that sustain AI and AI technical practices (even as AI may more overtly be co-narrating events in such practices through the use of simulations and digital twins for resource and manufacturing optimization and so forth).

7.2 Expanding narratives

Understanding AI as a process and narrative necessarily requires that we understand AI more broadly than might be usual or intuitive and requires expanding the field of practices that we consider when thinking of AI's coming to be. This is especially important if we would like to understand how AI can be 'ethical'—if the practices that sustain it (in its construction and maintenance) do not effectively aim at the good life or adhere to moral norms, then other interconnected practices (direct use cases) cannot fairly be said to be ethical where they rely on the perpetuation of harm through unethical practices.[6] To be responsible, the practices should be transparent. AI is a plurality of diverse interconnected practices, each of which must be understood to inform and activate responsibility. However, AI is also a plurality of narratives and persons' experience of it, as Coeckelbergh (2021b, 1632) notes, stories are always related to other stories, and persons whose lives are co-narrated by AI and who shape AI themselves. AI is a narrative field, as Haraway (1984) might say, and as a narrative field it is a space of political struggle and contestation.

AI can be a lived and embodied experience—it is enacted and experienced through bodies, situated through time and space, across a variety of locations with their own socio-cultural contexts. Narratives embed perspectives, whether they are the narratives of lived experience, or narratives about something such as imagined futures. To make sense of AI and AI practices, we must carefully search for the stories of all stakeholders in those practices (including and especially the non-obvious practices such as extractive ones). Stories capture voices, and bodies across different socio-cultural, localized contexts throughout which technical practices may take on different shapes and forms as technologies are appropriated differently, or affect people in different ways. Narratives can hold the rich details of the lived experiences of diverse stakeholders—narratives can draw light upon situated knowledges, upon the plurality of perspectives of different phenomena. Narratives are a foundation of transparency of diverse standpoints, without which we are left with so-called God tricks and views from nowhere (see Haraway 1988; D`ignazio and Klein 2020). This also corresponds somewhat with Ihde's (1990, 42) relational perspective on situated seeing, that is, there is no simple seeing, only seeing as a particular agent from a particular perspective or positionality.

To understand how AI comes to be and how it is appropriated across different practices, how it helps or harms persons, and how we might take responsibility and try to reform AI practices, we need to read the narratives of the underpaid geographically dispersed human labour that goes into training models, those of miners extracting mineral resources, persons subject to algorithmic judgments about the future of their criminal rehabilitation (Angwin et al. 2016), who might be misgendered at airports by security agents or technologies that only view persons in binaries (D`ignazio and Klein 2020), or of those simply about how AI helps individuals organise their lives or make decisions for better or worse (nudging algorithms or ChatGPT for example). The plurality of first-order and second-order stories across location and culture are intrinsically important for us all to come into contact with each other—responsibility is about understanding and negotiating the plurality of demands of different actors (Williams 2008), and justice can come in meeting them. The plurality of narratives about AI, about people's experiences within practice and process, illustrate the things that people from different socio-cultural backgrounds value, and where tensions between practice and beliefs about the good life arise. In understanding tensions endemic in practice across cultures and between stakeholders in different instantiations of a practice, dialogue can begin about how to accommodate visions of the good life for all, to truly enable us to live well, with and for others, *with AI and technology*,

[6] More detailed discussion along these lines can be found in (Krzykawski and Lindberg 2021).

in just institutions. Through recognising the voices of others we can practice ethics through *solicitude* (see Ricoeur 1994)—and also by understanding the plurality of experiences and competing or conflicting claims or demands in technological practice we can attempt to adjudicate on them through just institutions that strive (in principle) to see justice for all.

When a single dominant narrative is that which shapes AI, which then co-narrates the lives and experiences of others, an arguably evil abstraction can occur, what some might deem artificial stupidity, as its totalization proceeds in ignorance of diverse local contexts which it transforms (Fitzpatrick and Kelleher 2018; Krzykawski and Lindberg 2021). Such a situation is to be mitigated and can only be so where a multiplicity of narratives can co-exist (where the narrative field is adjusted), rather than be imposed from above. If AI is narrative and process, we must work to ensure that the master narrative is one that fairly reflects the tapestry of narratives it encompasses, that the master narrative does not abstract from local particularities, that AI is not artificial stupidity that inhibits our normal development as humans in specific milieux or limits our individual and collective capacities for ethical action in the world (Fitzpatrick and Kelleher 2018; Krzykawski and Lindberg 2021).

The lifeworld is a tapestry of narratives, and a tapestry of narratives of those involved in technical practices moving towards their visions of the good life, or being stopped on their journey towards the good life by technical practices that are not designed to recognise their rights and dignity. The recognition of diverse positionality and perspectives accommodated by narrative is in line with the kind of feminist epistemology that Donna Haraway (1988) has defended, an epistemology of dialogue and difference in meaning—narratives provide pathways to understanding different accounts of the world, accounts which we need to know to live together, and accounts which make difference and ethical tensions transparent and call us to responsibility. Narratives as such make the difference transparent. Narratives make the experience of the other (the other self), accessible, available, and understandable, and they are rich with detail relevant to the ethical inquiry of what it is to live well, with and for others, in just institutions. Narratives can help bridge the epistemic asymmetry between AI designers and those who are in some way affected by AI-driven outcomes (Nascimento 2019).

We must pay attention to expanded second-order narratives. Whilst it is important to pay close attention to those who are affected by or practitioners of particular technical practices, across locations and socio-cultural backgrounds, it is also important to pay attention to second-order narratives about technologies from diverse communities. In keeping with the arguments of Haraway, the perspectives of others on alternate or future states of affairs, or the dreams and aspirations or visions of AI, can fall within legitimate and diverse situated knowledges informed by local socio-cultural backgrounds and lived experience. The narratives of non-dominant groups need to be sought and promoted as alternatives to the narratives about AI which are championed by dominant groups and hegemonic powers. Such narratives, be they anti-colonial or otherwise, may function as antidotes or stabilisers to the neoliberal or libertarian narratives that are arguably currently largely influencing the shape and direction of technological evolution, practice, and the becoming of global communities in what have been fraught and dangerous states of affairs if we consider the consequences of AI in the form of disinformation campaigns that correlate with COVID and vaccine scepticism/denial and international civil and political strife. As we have argued, it is a narrative responsibility to forge such second-order narratives, and a moral responsibility to make them so. It is also a moral responsibility to seek these narratives as the effort to reach out, to respond to the capacities and dignity of others including in their desires for a better future. This is characteristic of the responsible agent who must negotiate the pluralities of moral patients and their rights and claims.

As we already stated, the interpretive process is open-ended and many possible interpretations of narratives too are possible, including interpretations of choiceworthy futures based on the values held by those doing the interpreting, including their visions of the good life as shaped by aspects of their cultures, and socio-political and economic situations and other elements of their upbringing. Such conflicts have been studied in hermeneutics and in relation to literature (Ricoeur 2007), and as in literature where interpretations can turn possibilities into actualities (Armstrong 1983, 341), in life different interpretations of the good life are contested spaces that can eventually militate into the actualisation of specific (again dominant) visions of the future. Ricoeur (2007) himself has worked extensively on the problem of the conflict of interpretations, which can help answer some questions in this domain, about valid visions of the future and of interpreting technologies and practices now, and application of this work can be a fruitful avenue of further research to compliment the insights presented here.

In the meantime, we can still tentatively posit on how choiceworthy narratives about the future may both emerge and be engaged with. Narratives can be seeded and explored in many spaces, starting with the individual (imagine the science fiction writer), to communities and clubs, any spaces that promote dialogue on present experiences and aspirations for the future, to spaces of public and international governance that foster both debate and action. Perhaps most critically, however, is the space of education (principally at all levels) as one that can promote dialogue between people (from potentially different socio-cultural backgrounds) on the future, both by building new narratives about futures that are more or less choiceworthy, and by engaging with

the narratives of others and helping to synthesise new ones. An excellent example of the possibilities of engaging with and generating narratives (and ethically salient ones at that) in the education context is Emanuelle Burton *et al's*. (2023) textbook, *Computing and Technology Ethics: Engaging through Science Fiction*. This book is intended for the ethical education of computing professionals, but in a world where everyone is a stakeholder in the future of computation (especially AI), everyone should have exposure to and a say in the topics dealt with in this book (its appeal may be more universal than its authors are aware). The book aims at supporting computing professionals in understanding the relevance of the plurality of values held by different people in society (depending on their particular standpoints) and bolstering their ability to practice ethical decision-making (Burton et al. 2023). These standpoints are reflected in ethically charged stories that express the vulnerability and concerns of their diverse characters (Burton et al. 2023). This textbook (and ultimately the syllabus promoted) presents stories that carry the multiple standpoints of their writers in the mode of science fiction, of different (often in the case, not particularly choiceworthy) futures, and stories that are intended to inform our views on the present (Burton et al. 2023, 18). In the classroom, again the conflict of interpretations may manifest itself and here divergent readings become an opportunity to learn and respect the other, indeed even as an other, and to develop moral and narrative imagination (Nussbaum 1998; von Wright 2002; Burton et al. 2023). This imagination is fostered through and supports identification with the other (at an appropriate distance), through empathy and compassion—it figures them as real and draws our attention to what is important through emotional engagement, and essentially, by rendering their figure, allows us to engage in solicitude, which is an affective, attentive and caring regard (Nussbaum 1985, 1991, 2003; von Wright 2002; Ricoeur 1994; Roeser 2011, 2017). Engagement with the text in an educational setting, especially one which uses ethical theory and science-fiction to compliment each other, has the possibility of developing minds that can work more collaboratively towards a future worth having, and there is potentially also the possibility for the reading of texts with themes of decolonisation that challenge dominant, entrenched and destructive narratives. Building the moral and narrative imagination may not just help develop more solicitous ethical decision-makers today, but also may strengthen their capacities to conceive futures worth having that can be attractive for the plurality. Narratives generate expectations, and the challenge remains keeping a closed gap between expectations and the (choiceworthy) realities that unfold. Education is where imagination, knowledge, and skill begin to form and blossom, and forms the earliest foundations of the ideas that feed into effective and fair regulation and governance.[7]

8 Conclusion

First and second-order narratives are an enlightening source of information that can contribute to significant states of transparency about process and practice—they illustrate and organise people (including their motivations and goals), things, and events into coherent wholes that are available for the kind of inspection that can reveal the answers to ethically salient questions. Narratives facilitate moral transparency, they render problems and tensions bare and call upon responsible agents to consider them, weigh them, and act—at least in dialogue with the patients of their actions. Responsibility becomes a matter of ensuring that lived narratives are choiceworthy, and that technologies that factor into them by means of practice co-author their stories in ways that respect the ethical intention of living well, with and for others, in just institutions. Narratives, when investigated, create transparency about the way things are, and in our case here in particular, how people and AI interact and co-shape each other through broadly understood AI practices, through cycles of processes and narratives in continuous, ambiguous, becoming. Whilst narratives shed light, they also inform and structure relations between people and things (human-technology-world relations and so forth), and therefore second-order narratives have an incredible potential to direct the evolution of technical and AI practices. The stories we tell about AI have the potential to set the course of its future development and appropriation. To this extent, these stories may even render the future partially transparent. Importantly this also means that the stories that are told by dominant groups better positioned to control narratives ensure that their power, including to shape the future, will be entrenched and so it is of incredible importance that systematically excluded groups and an overall plurality of people and communities have their voices and stories heard so that a more inclusive and democratic future of AI that figures into a plurality of visions of the good life can be shaped.

Acknowledgements This research was partly funded by the ADAPT Centre which is funded under the SFI Research Centres Programme (Grant 13/RC/2106_P2) and is co-funded under the European Regional Development Funds. This research was partly funded by EC-funded H2020 SwafS Project EUt+ EXTRAS (#101035812), and HE MSCA SE project EpisTeaM (#101129655). The authors are grateful to the reviewers for their insightful feedback.

Curmudgeon corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to

[7] For example see (Kerr, Barry, and Kelleher 2020).

wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Funding Open Access funding provided by the IReL Consortium.

Data availability There is no further data in support of this research.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ananny M, Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc 20(3):973–989. https://doi.org/10.1177/1461444816676645

Angwin J, Larson J, Matu S, Kirchner L (2016) Machine Bias, ProPublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: 19 October 2018)

Aristotle (2009) The Nicomachean Ethics. New Edition. Oxford: Oxford University Press (Oxford World's Classics)

Armstrong PB (1983) The conflict of interpretations and the limits of pluralism. PMLA 98(3):341–352. https://doi.org/10.2307/462275

Bringsjord S, Govindarajulu, NS (2022) 'Artificial Intelligence', in E.N. Zalta and U. Nodelman (eds) The Stanford Encyclopedia of Philosophy. Fall 2022. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/ (Accessed: 1 February 2023)

Burton E, Goldsmith J, Mattei N, Siler C, Swiatek S-J (2023) Computing and technology ethics: engaging through science fiction. MIT Press, Cambridge, Massachusetts

Capurro R (2010) Digital hermeneutics: an outline. AI Soc 25(1):35–42. https://doi.org/10.1007/s00146-009-0255-9

Cave E, Dihal K, Dillon S (ed) (2020) AI narratives: a history of imaginative thinking about intelligent machines. Oxford University Press, Oxford, New York

Chen J-Y (2015) Virtue and the scientist: using virtue ethics to examine science's ethical and moral challenges. Sci Eng Ethics 21(1):75–94. https://doi.org/10.1007/s11948-014-9522-3

Coeckelbergh M (2020) artificial intelligence, responsibility attribution, and a relational justification of explainability. Sci Eng Ethics 26(4):2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Coeckelbergh M (2021a) 'Narrative responsibility and artificial intelligence', *AI & SOCIETY* [Preprint]. Available at: https://doi.org/10.1007/s00146-021-01375-x

Coeckelbergh M (2021b) Time machines: artificial intelligence, process, and narrative. Philos Technol 34(4):1623–1638. https://doi.org/10.1007/s13347-021-00479-y

Crawford K (2022) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven London

D`ignazio C, Klein LF (2020) Data Feminism. Cambridge, MA: MIT Press

Donnelly BL, Siegel T, Lang MB, Siegel T, Donnelly M (2023) 'Showstopper! Strikes Plunge Hollywood Into Chaos With Pricey Movie Delays, Pay Battles and AI Anxiety', Variety, 19 July. Available at: https://variety.com/2023/film/features/hollywood-chaos-strikes-movie-delays-pay-battles-a-i-anxiety-1235673080/ (Accessed: 20 July 2023)

Fitzpatrick N, Kelleher J (2018) On the exactitude of big data: La Bêtise and artificial intelligence. La Deluziana. https://doi.org/10.21427/dfw8-m918

Fitzpatrick N (2021) 'Will the real quantified self please stand up?', In W. Reijers, A. Romele, and M. Coeckelbergh (eds) Interpreting Technology: Ricoeur on Questions Concerning Ethics and Philosophy of Technology. Lanham: Rowman & Littlefield Publishers

Foucault M (1980) Power/knowledge: selected interviews and other writings, 1972–1977. Random House USA Inc, New York

Haraway DJ (1984) Primatology is politics by other means. PSA Proc Bien Meet Philos Sci Assoc 1984:489–524

Haraway D (1988) Situated knowledges: the science question in feminism and the privilege of partial perspective. Fem Stud 14(3):575–599. https://doi.org/10.2307/3178066

Haraway D (1986) 'Primatology is Politics by Other Means', in R. Bleier (ed.) Feminist Approaches to Science. The Athene Series. Pergamon Press, Maxwell House, Fairview Park, Elmsford, NY 10523 ($12

Hayes P (2020) An ethical intuitionist account of transparency of algorithms and its gradations. Bus Res 13(3):849–874. https://doi.org/10.1007/s40685-020-00138-6

Hayes P, van de Poel I, Steen M (2020) Algorithms and values in justice and security. AI Soc 35(3):533–555. https://doi.org/10.1007/s00146-019-00932-9

Hayes P, van de Poel I, Steen M (2023) Moral transparency of and concerning algorithmic tools. AI Ethics 3(2):585–600. https://doi.org/10.1007/s43681-022-00190-4

Heidegger M (2013) The question concerning technology: and other essays. Reissue edition. New York; London Toronto: Harper Perennial

Hickok M (2021) Lessons learned from AI ethics principles for future actions. AI Ethics 1(1):41–47. https://doi.org/10.1007/s43681-020-00008-1

Ihde D (1990) Technology and the lifeworld: from garden to earth. Indiana University Press, Bloomington

Kaplan DM (2006) Paul Ricoeur and the philosophy of technology. J French Francoph Philos 16(1):42–56. https://doi.org/10.5195/jffp.2006.182

Kelleher JD (2019) Deep learning, Illustrated. MIT Press, Cambridge, Massachusetts

Kerr A, Barry M, Kelleher JD (2020) Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. Big Data Soc 7(1):2053951720915939. https://doi.org/10.1177/2053951720915939

Krzykawski M, Lindberg S (2021) 'Ēthos and technology', In: B. Stiegler (ed.) Bifurcate: There is No Alternative. Open Humanities Press, pp. 195–219

Kudina O (2021) "Alexa, who am I?": voice assistants and hermeneutic lemniscate as the technologically mediated

sense-making. Human Stud 44(2):233–253. https://doi.org/10.1007/s10746-021-09572-9
Latour B (2014) '"What's the story?" Organizing as a mode of existence', in J.-H. Passoth, B. Peuker, and M. Schillmeier (eds) Agency without Actors? New Approaches to Collective Action. London: Routledge
Future Learn (n.d.) (2024) 'What can we learn from Latour?', FutureLearn. Available at: https://www.futurelearn.com/info/blog (Accessed: 13 January 2024)
Loi M, Nappo F, Viganò E (2023) How i would have been differently treated. discrimination through the lens of counterfactual fairness. Res Pub (liverpool, England) 29(2):185–211. https://doi.org/10.1007/s11158-023-09586-3
Louridas P (2020) Algorithms. MIT Press, Cambridge, Massachusetts
MacIntyre A (2013) After virtue, Reprint. Bloomsbury Academic, London
Miller T (2017) 'Explanation in Artificial Intelligence: Insights from the Social Sciences', arXiv:1706.07269 *[cs]* [Preprint]. Available at: http://arxiv.org/abs/1706.07269 (Accessed: 22 May 2019)
Milojević I, Inayatullah S (2015) Narrative foresight. Futures 73:151–162. https://doi.org/10.1016/j.futures.2015.08.007
Nascimento F (2019) Technologies, narratives, and practical wisdom. Études Ricoeuriennes/ricoeur Stud 10(2):21–35. https://doi.org/10.5195/errs.2019.481
Nightingale E (2023) 'UK actors and Equity are battling the rise of AI-driven deepfake mods', *Eurogamer*, 20 July. Available at: https://www.eurogamer.net/uk-actors-and-equity-are-battling-the-rise-of-ai-driven-deepfake-mods (Accessed: 20 July 2023)
Nussbaum MC (1985) "Finely aware and richly responsible": moral attention and the moral task of literature. J Philos 82(10):516–529. https://doi.org/10.2307/2026358
Nussbaum MC (1991) The literary imagination in public life. New Lit Hist 22(4):877–910. https://doi.org/10.2307/469070
Nussbaum MC (1998) Cultivating humanity: classical defense of reform in liberal education: a classical defense of reform in liberal education, New Ed. Harvard University Press, Cambridge, Mass.
Nussbaum MC (2003) Upheavals of thought: the intelligence of emotions, 1st edn. Cambridge University Press, Cambridge
Reijers W, Coeckelbergh M (2020) Narrative and technology ethics. In: Reijers W, Coeckelbergh M (eds) Introduction. Springer International Publishing, Cham, pp 1–24
Reijers W, Gordijn B (2019) Moving from value sensitive design to virtuous practice design. J InformCommun Ethics Soc 17(2):196–209. https://doi.org/10.1108/JICES-10-2018-0080
Reijers W, Romele A, Coeckelbergh M (eds) (2021) Interpreting technology: ricoeur on questions concerning ethics and philosophy of technology. Rowman & Littlefield Publishers, Lanham
Richardson R, Schultz J, Crawford K (2019) Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=3333423 (Accessed: 28 February 2019)
Ricoeur, P. (1990a) *Time and Narrative, Volume 1: v. 1*. New edition. Translated by K. McLaughlin and D. Pellauer. Chicago, Ill.: University of Chicago Press
Ricoeur P (1990b) Time and Narrative, Volume 2. New edition. Translated by K. McLaughlin and D. Pellauer. Chicago, Ill.: University of Chicago Press
Ricoeur P (1990c) Time and Narrative, Volume 3: v. 3. New edition. Translated by K. Blamey and D. Pellauer. Chicago: University of Chicago Press
Ricoeur P (1994) Oneself as Another. Translated by K. Blamey. Chicago, IL: University of Chicago Press. Available at: https://press.uchicago.edu/ucp/books/book/chicago/O/bo3647498.html (Accessed: 24 October 2022)
Ricoeur P (2007) The Conflict of Interpretations: Essays in Hermeneutics. Edited by A. Steinbock and D. Ihde. Evanston, Ill.: Northwestern University Press
Roeser S (2011) Moral emotions and intuitions. Palgrave Macmillan UK. https://doi.org/10.1057/9780230302457
Roeser S (2017) Risk, technology, and moral emotions, 1st edn. Routledge, New York
Romele A, Severo M, Furia P (2020) Digital hermeneutics: from interpreting with machines to interpretational machines. AI Soc 35(1):73–86. https://doi.org/10.1007/s00146-018-0856-2
Russell S, Norvig P (2016) Artificial Intelligence: A Modern Approach, Global Edition. 3rd edition. Boston Columbus Indianapolis New York San Francisco Upper Saddle River Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto Delhi Mexico City Sao Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo: Pearson
Sartori L, Theodorou A (2022) A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. Ethics Inform Technol 24(1):4. https://doi.org/10.1007/s10676-022-09624-3
Schwartz O (2018) '"The discourse is unhinged": how the media gets AI alarmingly wrong', *The Guardian*, 25 July. Available at: https://www.theguardian.com/technology/2018/jul/25/ai-artificial-intelligence-social-media-bots-wrong (Accessed: 20 July 2023)
Simondon G (2017) On the Mode of Existence of Technical Objects. Translated by C. Malaspina and J. Rogove. MinneapolisUniversity of Minnesota Press: Univ Of Minnesota Press.
Stern S (2023) *Don't be deluded by the exaggerated claims made for AI*. Available at: https://www.ft.com/content/dc7d6217-a7ba-451a-a18f-d55356c7fae3 (Accessed: 20 July 2023)
Stiegler B (2013) What makes life worth living: on pharmacology, 1st edn. Polity, Cambridge, UK
Stiegler B (1998) Technics and Time, 1: The Fault of Epimetheus, 1st edn. Translated by R. Beardsworth and G. Collins. Stanford University Press, Stanford, Calif
Stiegler B (2010) For a new critique of political economy, 1st edn. Polity, Cambridge; Malden, MA
Tu, Y.-C., Thomborson, C. and Tempero, E. (2011) 'Illusions and Perceptions of Transparency in Software Engineering', in 2011 18th Asia-Pacific Software Engineering Conference. 2011 18th Asia-Pacific Software Engineering Conference, pp. 365–372. https://doi.org/10.1109/APSEC.2011.42.
Tu Y-C. (2014) Transparency in Software Engineering. Thesis. ResearchSpace@Auckland. Available at: https://researchspace.auckland.ac.nz/handle/2292/22092 (Accessed: 19 October 2018)
Turilli M, Floridi L (2009) The ethics of information transparency. Ethics Inform Technol 11:105–112
Vallor S (2018) Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Reprint edition. Oxford University Press.
van de Poel I (2011) The relation between forward-looking and backward-looking responsibility. In: Vincent NA, van de Poel I, van den Hoven J (eds) Moral responsibility: beyond free will and determinism. Springer, Netherlands, Dordrecht, pp 37–52
van de Poel I, Royakkers L, Zwart SD (2015) Moral responsibility and the problem of many hands, 1st edn. Routledge, New York
Verbeek PP (2015) Cover story: beyond Interaction: a short introduction to mediation theory. Interactions (ACM) 22(3):26–31. https://doi.org/10.1145/2751314
Verbeek PP (2016) Toward a theory of technological mediation a program for postphenomenological research. In: Friis JKBO, Crease RC (eds) Technoscience and postphenomenology: the Manhattan papers. Lexington Books, London, pp 189–204
von Wright M (2002) Narrative imagination and taking the perspective of others. Stud Philos Educ 21(4):407–416. https://doi.org/10.1023/A:1019886409596

Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol 31(2):841–887

Williams G (2008) Responsibility as a virtue. Ethical Theory Moral Pract 11(4):455–470

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.