Grant Agreement Number: 824671

SUPER MoRRI – Scientific understanding and provision of an enhanced and robust monitoring system for RRI

D 5.1: Case study co-creation methodology report

(How) Can you build morality into artificially intelligent systems?

Author(s): Kjetil Rommetveit, Ingrid Foss Ballo
Submission Date: 14.06.2023
Version: 2
Type: Research protocol and case project report
Dissemination Level: public
This deliverable is an early version of an article to be published in a scientific journal.

Project website: www.supermorri.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824671.The opinions expressed in this document reflect only the authors' view and in no way reflect the European Commission's opinions. The European Commission is not responsible for any use that may be made of the information it contains.

Introduction

As more and more technological systems become 'intelligent' and autonomously operating actors enabled by machine learning, robotics and artificial intelligence (AI), doubts about the ethical and societal viability of such technologies come to the fore of public policies. According to Eurobarometer, Europeans are increasingly concerned about impacts of AI, especially automation of tasks at work (Eurobarometer 2021), and specific websites dedicated to mapping of AI incidents have become popular[1]. Initial debates were predominantly infused by science fiction and voiced by prominent people in science and technology: fears of super-intelligent machines ('general AI') gaining autonomy and turning against its human users and developers, even eradicating the human race (Boström 2002). Yet, as AIs to some extent become reality, debates are about by more mundane issues, such as social sorting (Pasquale 2015), 'bias', power imbalances and justice, election manipulation (Van Dijk 2021), ubiquitous surveillance (Zuboff 2019), privacy infringements (Rommetveit and van Dijk 2022) and accidents provoked by self-driving cars. One immediate response to this state of affairs has been a flourishing of ethical reports, analyses and meta-analyses, of the 'ethics of AI' (for an overview, see Fjeld et al. 2020)). In Europe, a prominent example of this is the so-called AI regulatory package, which includes law and soft law regulation, and foreshadowed by an ethics report by the High Level Expert Group on AI Ethics (hereafter AI HLEG).

A precursor to present practices based in values-in-design occurred in Asimov's three principles of robotics, introduced in science fiction as early as 1942 (Asimov 1942)[2]. In Asimov's story, these ethical principles are described as 'built most deeply into a robot's positronic brain' (Asimov, 1942). Yet, it would take several decades before any such possibility would start to be imagined and experimented with in actual robotic or digital systems (including so-called artificial intelligence), or by ethicists and philosophers. The ethics of ICTs were increasingly discussed and analysed in novel fields such as computer ethics (Moore 1985, Tavani 2007), engineering ethics and robo-ethics (Lin, Abney and Bekey 2012) during the 1980s and 90s. Yet, it was not until later, i.e. in the 1990s and 2000s, that prospects of actually hardcoding moral principles into a robots architecture became a topic, triggering imaginations and prospects of a new field, *machine ethics*, where moral principles can be designed and hardcoded into the robot's 'brain' (Allen, Varner and Zinser 2000, Allen and Wallach 2009). Robot morality and rights for artificial agents have been hotly debated in political arenas, such as the European Parliament's debate over 'electronic personhood' in 2016 (Rommetveit et al. 2020). It is also presaged in global documents and standards, such as the IEEE Ethically Aligned Design (IEEE 2017, 2019), and in the IEEE's 7000 Ethics for Autonomous and intelligent Systems series (IEEE 2023).

These regulatory developments are novel in a least two ways: first, by embracing a risk-based approach to values-based regulations, i.e. focusing on 'significant risk to the health and safety or fundamental rights of persons', (Trilateral Research 2022). The language of risk assessment and management is historically alien to the universe of moral and legal principles, but this is now changing (Van Dijk et al. 2016). Second, new approaches in values-based design will assess such risks, as well as ethical (and legal) principles, and try to take them into account at the earliest stages of designing AI systems (AI-HLEG 2019), rendering them in principle 'ethical, lawful and robust' (*Ibid.*) *by design*. Whereas these developments do have precursors,

[1] https://incidentdatabase.ai

[2] 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2) A robot must obey orders given it by human beings except where such orders would conflict with the First Law. 3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

i.e. in the field of values-based design, their introduction to governance and regulation is new. In the context of the AI-HLEG report, specific guidelines for ethics-by-design have been developed by a European Commission expert group (EC 2021), providing specificity for implementation of the principles outlined in the AI-HLEG report. This turn to risk- and design-based solutions also entails a shift in the sites of governance: towards engineering and infractructuring work. They render normative standards part of the very building blocks of technologies, infrastructures and markets, and so also displace the focus of governance. There is by now a plethora of standards oriented towards implementation, protection and enhancement of human-centric values, such as (in order of historical emergence) the (British) BS8611, the IEEE Ethically aligned Design initiative (IEEE 2016, 2018), followed up by the IEEE 7000s-series (including more than 10 'ethical standards', IEEE 2023), and the ISO ISO/IEC 38507 on Governance implications of the use of artificial intelligence by organizations.

In previous research on data protection we have identified this turn towards design-based regulation as inscribed into a techno-regulatory imaginary (Rommetveit and van Dijk 2022), whose defining characteristic is that the very problem articulation is now imagined to be situated inside the technologies and infrastructures themselves, partially outside of classical regulatory sites (government, parliament, state bureaucracies) oriented towards standardisation and infrastructuring. Their main mode(s) of implementation reside in exactly the risk- and the design-based approaches to management and regulation.

In this brief report we investigate these developments, pursuing the question of what happens to values and principles as they become matters of design and engineering. As stated in the Super-MORRI D5.1 the case study will investigate efforts (in governance and research laboratories) to build ethics and values into autonomous (digital) systems. It will focus on how ethics in design is being implemented in practice, where and by whom (p. 21). The aim is to document changes to policy practices as indicated by the increasing use of design-oriented language and practice. The policies addressed by our research are specifically relevant for the RRI keys 'Ethics' and 'Governance'. Our study can be situated within an RRI framework both through its emphasis on the product and process of AI technologies (Von Schomberg 2011), and through emphasis on mutual alignments of variously involved actors with regard to *reflexivity*, *responsiveness*, *anticipation* and *deliberation* (von Schomberg 2011, 2013; Stilgoe et al. 2012; Owen 2015; see also Callon et al. 2001; Guston 2014; RRI Tools 2014).

Period / Site	Google	Springer	Nature	IEEE
1990-95	5 370	554	4	40
1995-2000	7 800	762	6	76
2000-2005	14 200	1 022	6	150
2005-2010	17 400	1 842	11	256
2010-2015	18 400	3 647	43	301
2015-2020	27 600	11 975	566	447
2020-2023	136 000	26 911	1 343	474

Table 1: time-series for hits on search words 'artificial intelligence' AND 'ethics' AND 'design'. For the IEEE the search words were limited to 'ethics' and 'design'. From 2015 and onwards we see sharp intensifications of occurrences.

Table 1 demonstrates broad changes to the ways in which two RRI keys, ethics and governance, are talked about and practiced. In broad terms, these are changes to the fields of ethics and engineering, through which these are brought into closer contact and mediation

(Latour 1994, Verbeek 2006). The mediating factor in this case is the language of design, which holds out the promise to more closely align digital technologies with ethical values, such as autonomy, fairness, privacy and accountability (EC 2021). As analysed by mediation theory, the meaning of 'ethics' and 'values' can be expected to change as they become embedded within technical and material infrastructures, and with (software and hardware) engineering practices. And, as already mentioned, this shift in the meaning (of ethics and governance) is inseparable from shifts in *site*, where practices of ethics and governance become implemented in more privatised, market-based and technology-centered sites.

Hence our research question: *what happens to rights and values as they become subject to design and engineering.* Because there is not (as we will show) one single answer to that question, we shall proceed in this text to lay out 5 design articulations, each of which describes the space of mediation (captured by the term 'design') differently. This does not mean that our design articulations are mutually exclusive: at least some of them are likely to co-exist and mutually fortify each other; others may come into conflict or be mutually exclusive. We first lay out our 5 design articulations, then comment on some such overall relations between them.

Design articulations

Our first two design articulations correspond with fairly classical positions, according to which (institutional and philosophical) ethics and engineering are different knowledge practices with distinct ways of knowing and different claims on truthfulness and veracity (Latour 2013). This implies that, when brought together, one of these tends to dominate or dictate the other involved parties and knowledge practices, in an asymmetrical relation (Gorman 2010).

1: Ethics rules

On the face of it, it seems clear exactly who knows about values and ethics, namely professional ethicists, many of whom have a background in philosophy. Accordingly, the task should be to articulate the right and fitting principles for AI systems, such as a recommendation system, a driverless vehicle, a face recognition system or a care robot. This way of configuring the problem and its design space begins and ends with ethics, insofar as the main emphasis is on the ethical analysis and its resolution in philosophical and ethical terms. Within philosophy this tendency towards solving problems in philosophical terms is strong, and this also has a strong impact on philosophical ethics. Yet, our category is somewhat broader, insofar as we include any position that implies that the problem is primarily a normative one, and that the normative question should be settled before (f.i. as a legal principle), and outside of, the task of designing and implementing a technical or organisational problem. Examples of this approach from our ethics corpus include (Bryson 2018, Ryan 2020, Liao et al. 2020).

The ethical problem, then, could be framed according to a number of ethical and philosophical positions, and also legal ones. In the AI ethics literature, typical positions are virtue ethics, Kantianism (duty ethics or de-ontological ethics) and utilitarianism. Other approaches based in philosophical ethics could be: care ethics, feminist ethics, meta-ethics, phenomenology and hermeneutics. Whereas there is great diversity, a possible common denominator would be the tendency to mainly use the AI ethical problem as a case in which philosophers and ethicists get to exercise their arguments, frequently grounded in long-standing debates between theoretical positions. As such, autonomous digital systems and AI do not pose radically new requirements but serve as occasion for re-casting philosophical

ethics in a new setting. That is, the problem framing is largely discipline-based, and not emerging from a context of application. Yet, the question then becomes: presuming that the ethical problem is dealt with: how is it to be translated into material and technological practice? Is this question dealt with, or is it treated as insignificant? Yet, commonly, this question is not answered within the frame of this articulation.

2: Engineers in the driver's seat

A second position becomes almost the opposite of the previous: here, the engineers take the lead. By defining the problem space as a task for engineering, the ensuing disciplinary spaces and intellectual relations of problem solving are also shaped. As one instantiation of this position, consider the statement by Prof. Ali Hassami at an open panel at the main privacy conference in Europe (CPDP), *Can ethics be standardised?* Hassami claimed that philosophers have had the chance of setting ethics on a firm basis since the times of Aristotle, but basically, they have gotten nowhere: ' Aristotle's virtue ethics is more than 2000 years old, still there are no standards for it'. Hassami did not refer to standards in an 'old' sense (i.e. as applying to people and to norms of proper behaviour, rather than to technical things, cf. Busch 2011); rather, he was referring to technical standards for engineering. Hassami was the chair for the IEEE's ethics standards series and has been vocal on the need to include ethics as part of engineering discipline and practice. A main example of this approach can be found in the IEEE standard 7007, where various ethics approaches (virtue ethics, utilitarianism, kantianism) are transformed into machine-readable formalised language. According to Hassami, the aim of this standard series is to '...offer a way of addressing ethical concerns to engineers and scientists who were involved in articulating, designing and developing autonomous intelligent systems…' (Bussemaker and Hassami 2022). This inclusion of ethics into engineering practice is necessitated by the evolution of autonomous intelligent systems. It entails that we 'throw in one other category (ethics) that has been eluding us, if you like, until AI came about'. The reason for this argument is that AI (and, more broadly, machine learning) enable more autonomous behaviours in machines, including simple decision-making capacities (ibid.).

The audience of this approach is 'engineers and developers' as it is they who will have to render ethics an object of engineering interventions. It is '…largely a technocratic approach' although 'the standard recommends consultation with key stakeholders' (ibid.) for inputs on key ethical and moral (even religious) matters. This design articulation therefore exists in a rather tense relation: between the requirements of engineers aiming to standardise language and render ethical considerations machine-readable, and the need for broader 'stakeholder involvement' and inclusion of a great variety of values and interests. It is reflected in how Hasssami refers to the need to keep an open mind to different ethical positions, including non-western ones, whereas when it comes down to the making of actually machine readable 'ethical ontologies', the alternatives are reduced to the classics: virtue ethics, de-ontology and utilitarianism. That this approach is not shared by all, or most, engineers, is clear: within the 7000 standards series it makes for an exception. It is specifically pursued however in IEEE 7007, *Ontological Standard for Ethically Driven Robotics and Automation Systems*, which seeks to formalise and operationalise virtue ethics, Kantian de-ontology, and utilitarianism. the operationalisation of ethical and moral theories may, for instance, look like this:

A Norm is a Method entity that describes a set of rules and methods governing behavior expected for norm-aware agents. Norm types are derived from respective ethical theories and are possibly influenced by agent social contexts. An ethical theory is a systematization of concepts specifying or recommending aspects of morally correct behavior based on philosophical values and the characterization of right and wrong conduct. For norm aware agents, normative ethical theory is concerned with the practical means of determining a moral course of action.

```
(forall (n) (if (Norm n) (ERAS-TLO:Method n)))

(forall (t) (if (EthicalTheory t) (ERAS-TLO:Method t)))

(forall (n) (if (Norm n)
          (exists (t s)
            (and (EthicalTheory t)
                 (SocialCollection s)
                 (specifies_norm_modality t n)
                 (is_prescribed_by n t)
                 (influences_norm_applicability s n)))))
```

(from: IEEE 7007, p. 25)

This approach seems to bypass the tricky question of whether this can actually be done, as the language of normative principles and values is very different from highly formalised language needed for machine readability ones[3]. In a well-known work on machine ethics (Allen and Wallach 2009), the authors proposed the idea of full moral agency for machines, but they later backtracked and argued the need to pay more attention to the practices and environments in which machines are developed and used. Experiments have also been carried out to implement simple ethical rules in robots (Vanderelst and Winfield 2018), concluding that ethics should not be built into machines. The decisive fact however is the *ethical simplicity* of the experiment: whereas highly sophisticated in technical terms, the experiment is still very far away from actual implementation in complex, unstructured environments, or from anything approaching the complexity of a real-world moral dilemma. It is highly dubitable, therefore, if this approach can be implemented in the sense of rendering machines as full moral agents. It may however serve other purposes, such as mobilising the engineering community by assigning to it its own proper engineering approach to ethics.

In the next two design articulations, we describe proponents of design that regard such limitations as fundamental, and therefore direct attention to the human designers, creators and operators of intelligent and 'autonomous' digital systems.

3: Educating the engineers

We have traced strong commitments, amongst policy makers and people implicated in AI ethics, for more human-centric approaches to AI ethics (AI-HLEG 2019), including in our literature review and amongst our interviewees. Most iterations of this commitment do not take the routes described above. Rather, they are human-centric, in the sense of focusing standards on the behaviours of human creators, designers and operators of AI systems. As stated by two interviewees, both with a strong foothold in philosophy:

[3] As stated by one expert in human-computer interactions: There is a difference between the moral reasoning linked to human rights and the attempt of solving an engineering problem, which is technically and mathematically specified (cited from: Rommetveit and Van Dijk 2022).

So my own view is we should not be trying to have more ethical technology, we should be trying to normalize and help people learn about processes that will have more ethical results (philosopher A).

what's really grounding my current projects ... it's really human wisdom about machines, not wisdom in machines, although I think the two issues are not completely independent (philosopher B).

This realisation comes in part from thinking about limitations to what both humans and machines can do, and from a general scepticism about anything approaching human-like artificial intelligence ('general AI'). Philosopher B described this as a learning process through which he came to focus less on the machines, and more on their socio-technical context and surrounding environments. It was described as being more realistic, and (critically) contrasted with science fiction-like scenarios and transhumanist imaginations, which nevertheless have strong impacts on public discourse: *I think people immediately want to jump to Bostrom-like scenarios, which say what if the AI decides that making paperclips is the-, you know, it's like no, that's not the issue here* (philosopher B).

Against such speculative positions on ethics, this philosopher claimed a need to focus on more everyday mundane tasks, and the consequences they may have on people and everyday relations. Here, the threat perception and 'the problem to be addressed' is more directed towards a thousand mundane operations and technologies, as this is where most impacts of AI will be, and are being, experienced. This problem was associated to the ways in which engineers and technologists are educated to not be aware of a thousand ethical decisions made during system's development, design and implementation:

They're saying this is what matters in the world, this is the thing we want our algorithm to be good at, and that is to say that some things are more valuable than others. And it's fine for them to do that, you have to make value choices in the process of design. The question then becomes which value choices, and now most recently where my work has gone to is how do we help computer scientists and technologists and roboticists, how do we help them to make those decisions in an intelligent way. They're going to have to make these ethical choices, these value-based choices, they have to do that in order to build the technology, so how do we help them do it better (philosopher A).

Hence, the need to teach ethics to engineers, and to include this in a design process in which more and better awareness of ethical issues can become part of engineers' everyday work tasks. This position makes good sense, as it addresses a *sine qua non* for the implementation of systems of all kinds, i.e. the engineers and software developers. However, it may also have come from a close vicinity of practices: it is after all much easier to move into the engineering department at your university, than into Google or other large AI-developing companies. For instance, the firing of Google ethicist Timnit Gibru was specifically mentioned as an example of such alienation and lack of agency relating to large companies. Yet, if you can influence engineers to think more ethically, they may bring this skill even into large powerful companies at later stages, and so this is one possible way of circumventing a lack of influence on main societal institutions and corporations.

Arguably, this position held stronger sway with our US respondents than the European ones. This was not necessarily due to any strong difference in opinion, but rather to differences of culture, economy and institutional arrangements. One (philosopher B) described the US environment as 'Wild West', another (philosopher A) described a highly fragmented environment with little or no coordination across actors, institutions and sectors. The argument was that, whereas there is great emphasis and urgency underpinning efforts to integrate ethics, AI and design, there is no concerted movement or organisation to take care of and coordinate such efforts, at least not in the US.

The perception was also prominent that these things are, in some ways, handled in better ways in Europe. Both philosophers (A and B) were highly aware that this approach, focusing more on what is going on in education, is insufficient, as significant developments take place in privatised centers of power. Here, standards-based solutions may hold some promise, as they (arguably) involve a greater set of actors (stakeholders), including private ones. Yet, these informants were sceptical of such efforts:

I guess I've just seen enough of these kinds of efforts, that they take 20 years and they end up with something that's not terribly useful. We need things right now, so let's put at least some kind of tools into people's hands so that they can start to make real progress now (philosopher A).

Philosopher B made reference to Europe, and European experiences with 'having a multiplicity of viewpoints represented', through user involvement and public engagements, but also more public regulation, amongst them lately 'multistakeholders' included in standardisation processes. Whereas this was partially explained by a lack of familiarity with those (regulatory, European) settings, the argument was also coupled to the kinds of expertise required for properly training and transmitting ethical skill and sensibility. He referenced experiences from engineering, computer and business ethics, where (gradually) teaching had been taken over by non-philosophers, and 'increasingly not by people who have been trained in philosophy departments'. This pointed to another boundary drawn around this articulation: it certainly came nowhere near the ivory tower approach in which ethical principles would be imposed top-down, which was also strongly criticised. This pointed to a difficulty, and relative novelty, within the community of academic philosophers and ethicists: on the one hand, ivory-tower philosophy is not sufficient; on the other, non-philosophers may not be capable of transmitting the

4: Educating the stakeholders

A next set of actors comes close to the previous design articulation by sharing many of the same concerns: the potentially harmful and de-humanising role of AI technologies on human relationships, and their threats to human dignity and autonomy. Also these informants articulate the means of addressing such threats through human-centric means focused on the human designers, operators and users rather than the machines themselves: *my opinion is that ethics is something that is proper to human beings, maybe some animals, but that's disputable, but to human beings. And ethics is a reflection, it's not something that you can really automate* (engineer C)

Perhaps because these were closer to European governance structures, and coming out of engineering and robotics, they did not however articulate the same skepticism towards standards and regulations. Indeed, roboethics has been an established part of European

innovation and governance since the early 2000s (Veruggio 2006, Rommetveit et al. 2020). Our informants described their encounters with AI ethics therefore as gradually emerging out of these developments, and as closely related to standards and regulations: *I mean there are lots of ethical aspects to standards. One which I didn't realise until I got heavily involved about ten or twelve years ago when I started to work in standards* (engineer A). All made reference to maturing technological capabilities, initially in robotics then gradually also encompassing AI and AI-like technologies and networks, through which machines take on more human-like, simple decision-making capabilities. This triggered ethical reflection and brought them into various ethics and governance boards and fora.

The description of a complex and unstructured technological and regulatory environment is also shared with the informants of the previous section. Much of this resides in the technology itself, which is much more distributed, networked and privatised than for instance bioethics regulatory efforts: *the body is kind of a natural safe for the actual DNA, and we (won't have) [0:13:46] the equivalent for the virtual DNA. So this is the difference the way I see it. And many people, they distribute their virtual DNA without being aware* (engineer B). This ubiquitous nature of data and information ('digital DNA') is one reason why efforts to tackle ethical problems should be addressed at infrastructural levels, as these are extremely encompassing and indeed in many cases the main enablers for the global spread of data and (autonomous) computation:

standards are a kind of invisible infrastructure of the modern world. Everything around us. You've only got to look around your room to see standards in everything. The quality of the mug that I'm drinking out. Obviously, the laptop, the Wi-Fi, the fact that we can speak to each other like this. None of that wouldn't happen without standards (engineer A).

Our engineering informants were not all located in Europe, but were closely aligned to various European governance initiatives, and also to the IEEE. They described an evolving regulatory landscape, beginning with efforts to regulate robotics. As already mentioned, many such initiatives were traced back to the early 2000s, but it would take almost 20 years for many of them to come to fruition, institutionally speaking:

Anyway, so we started this initiative, the launch was in April 2016. And the mission of this initiative that I was chairing, still chairing but it's going to be probably ending soon now, in 2016 was to raise awareness about ethical issues related to AI (engineer C).

This quote refers to the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, whose main outcome eventually became the *Ethically Aligned Design*, a written textual document elaborated in a bottom-up way by engineers, ethicists and other 'stakeholders' (IEEE 2016, 2019), with the latest version (of three) being officially adopted by the IEEE board. The drafting process of the text took close to 5 years and was conceived as bottom-up process open to anybody with an interest. This initiative fed into the IEEE 7000 standards series, with at least 12 (14?) different expert groups engaged in various topics dedicated to ethical alignment. As stated, most of these are not machine-centric but rather targeted at the human operators. Whereas engineers remain important, at this point much larger groups ('stakeholders') come into view, as standards require commitment from groups as various as government regulators, business corporations, technology developers and vendors, and (to some extent) civil society. Processes of standardisation may also be linked to more downstream initiatives, such as certification schemes (based on the standards), and to

efforts to 'educate the public', although how that is supposed to happen was not clear from our interviews (nor from the document study).

Finally, the work of the AIH-LEG expert group in many ways superseded and built on this work of standardisation:

In 2018 the European Commission has wanted to build their AI plan, every country was doing this of course, you know why, and the Commission to create its own AI plan put together this high-level expert group of 52 members. And I applied and I was selected, so this is how IEEE and the high-level expert group in a way conflated together. And we produced as you know the ethics guidelines and the policy recommendations as well within the high-level expert group.

There is therefore a history to be told here, about the emergence of this articulation: it starts with early (European) roboethics initiatives (Veruggio 2006), passes through the IEEE Ethically Aligned Design, then goes onto the IEEE 7000 series, and culminates (for now) with the AI High Level Expert Group.

5: AI transformation?

Our fifth (and final) design articulation is possibly more radical than the previous four, insofar as it seems to take the mediation of 'ethics' with AI more seriously. It does so by blurring the distinctions between human and artificial intelligence, yet such blurring may take place in very different ways and based on highly differing accounts of both human and technological agency. There are two versions here, one that believes in 'strong AI', that is a general artificial intelligence, both as threat and as possibility (we have seen that this possibility is largely rejected, at least by the previous two articulations; not as in principle impossible, but rather as unattainable based on the present state of art). One case in point was articulated as follows, by an engineer that works for a well-known civil society organisation:

I would say it's the social implications of the prospects of the technology, because once we get past a certain level there are implications that... you know, it's pretty clear that things would change qualitatively and that society is just not prepared for that.

Next, there is a weak AI version, one that does not believe in a radically transformative potential, at least not in the sense conveyed by the previous informant: *I don't think general AI is a very sensible or a well-theorized concept*. This informant (engineer D), who works for an organisation that carries out fundamental research into 'intelligence' in its various aspects, nevertheless finds the potentials of merging human and machine intelligence:

We found ourselves more interested in the capacity of AI to foster, support human morality, largely seeing it as a conflict between what's sometimes called the paradox of automation, which is as you use a tool you get less good at doing things without it. So to the extent that AI could support human morality, it might also tend to supplant moral decision making. So the question that we were focusing on, continue to focus on really, is how AI tools, modern technological tools can strengthen our human capacities rather than replace them.

As an example, a robot or an AI could be used in therapeutic situations, to support therapists in work considered repetitive and tedious. Or, an AI could be used to sift through large amounts of (big) data on a certain moral dilemma (he mentions kidney transplants), to come

up with alternative suggestions not immediately visible to human decision makers operating through an ordinary qualitative understanding of the problem at hand. This articulation, then, is similar to the 'strong AI' version in that it considers moral and technological change simultaneously. However, it is much more skeptical (or: scientifically experimental) on behalf of what AI can actually deliver.

Discussion and conclusion

The 5 AI design articulations demonstrate in and by themselves moral, or techno-moral change: the introduction of morality and ethics to the world of standards and engineering is shifting the meaning of morality and ethics. On the standard account ('ethics rules') the philosophers and ethicists are still in charge, and the question remains how to articulate the right reasons and best arguments. However, this is only one out of 5 articulations: the next one, where the engineers are taking the driver's seat, poses decisive limitations on the articulation of ethics. First, ethics is delimited to (mainly) rules-based accounts (deontology and utilitarianism), with a possible minor exception for virtue ethics (where 'morality' emerges in a somewhat more bottom-up fashion, drawing on machine learning techniques. Articulation number 3 may be seen as taking us back to a more classical, human-centric, notion of engineering ethics, as it becomes a matter for the ethicists (or similar) to educate the designers and programmers of AIs, preferably in engineering school. Whereas similar in philosophical orientation (i.e. pragmatism), articulation No. 4 moves further into the landscape of standards, although seeing them more as foci for human, organisational, business and engineering efforts. Yet, the meaning of a 'right' or a moral principle can here be expected (and observed) to be shifting towards meanings accessible to all, not any longer the exclusive property of ethicists. Articulation 5 mover even further from a classical understanding of ethics, as here it becomes a matter for human and machine intelligence as a joint, possibly enhanced, enterprise.

None of this however implies, by necessity, actual socio-technical change: ethics is 'soft law', and may not come with hard effects although that has certainly been argued with force (Tallacchini 2009). Yet, what it does signify, is an increasing emphasis on (digital, AI) infrastructures as centers for governance: it signifies therefore a partial movement out of traditional institutions (governance), or rather their possible extension and merger with digital infrastructures as increasingly central to a number of important and critical functions. The main sites of this shift are the standards organisations and committees. This shift can be seen in law (a more reliable indicator than ethics due to its important societal and regulatory function), and in increased emphasis on (digital) infrastructures in science and technology studies, data studies and governance studies. This is sufficient for outlining a 'theory of change' in the Super-MORRI project: prior to the emergence of machine ethics and ethical standards such as the iEEE 7007, 'ethics' was always human-centric in the specific sense of targeting the professional (human being) making or using a technological artefact.

References

IEEE. (2018). *Ethically aligned design. A vision for prioritizing human well-being with autonomous and intelligent systems.* The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence, 12*(3), 251–261.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong.* Oxford University Press.

Asimov, I. (1950). "Runaround". *I, Robot* (The Isaac Asimov Collection ed.). New York City: Doubleday.

Busch, L. (2011) Standards: Recipes for Reality. MIT Press.

Bostrom, N. (2002). "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." Journal of Evolution and Technology, 9. http://www.nickbostrom.com/existential/risks.html

Bryson, J.J. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics Inf Technol* 20, 15–26 (2018).

Callon, M., P. Lascumes and Y. Barthe (2001), Agir dans un monde incertain. Essai sur la démocratie technique, Paris, Le Seuil.

European Commission (2021) Ethics By Design and Ethics of Use Approaches for Artificial Intelligence. Version 1.0 25 November 2021. Brussels.

Eurobarometer (2021) European citizens' knowledge and attitudes towards science and technology. September 2021. Accessed on June 13. from: https://europa.eu/eurobarometer/surveys/detail/2237

Fjeld, J. Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI." Berkman Klein Center for Internet & Society, 2020.

Gorman, M. E. (Ed.). (2010). *Trading Zones and Interactional Expertise: Creating New Kinds of Collaboration*. The MIT Press. http://www.jstor.org/stable/j.ctt5hhhrw

Guston, D. (2014), 'Understanding "anticipatory governance"', *Social Studies of Science, 44 (2), 218–42.*

Hessami, A. and Bussemaker, F. (2022) Accountability on the Digital Age. Accessed on: June 13., 2023 at: Videohttps://i4ada.org/dialogues/prof-ali-hessami/.

IEEE. (2016). Ethically aligned design. A vision for prioritizing human well-being with autonomous and intelligent systems. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

IEEE. (2018). Ethically aligned design. Version 2.0. A vision for prioritizing human well-being with autonomous and intelligent systems. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems.

IEEE (2023) The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 7000 series. Accessed on June 13th. at: https://standards.ieee.org/industry-connections/ec/autonomous-systems/

Latour, B. (1994) On Technical Mediation. *Common Knowledge,* FaIl 1994 V+3 N2.

Latour, B. (2013) *An Inquiry into Modes of Existence: An Anthropology of the Moderns,* Catherine Porter (tr.), Harvard University Press, 2013.

Liao, S. Matthew (ed.), *Ethics of Artificial Intelligence* (New York, 2020; online edn, Oxford Academic, 22 Oct. 2020).

Moor, J. H. (1985) What is Computer Ethics? *Metaphilosophy*, 16(4).

Owen, R. (2015), 'Responsible research and innovation: options for research and innovation policy in the EU', accessed 14 April 2023 at http://goo.gl/DMCyNZ.

Rommetveit, K., van Dijk, N., Gunnarsdottír, K. (2020) Make way for the robots! Human- and machine-centricity in constituting a European Public-Private Partnership. *Minerva: A Review of Science, Learning and Policy.* 58(1):47-69.

Rommetveit, K., Van Dijk, N. (2022) Privacy Engineering and the Techno-regulatory Imaginary. *Social Studies of Science*.

RRI Tools (2014) 'RRI Tools: towards RRI in action', accessed 14 June 2023 at https://rri-tools.eu

Ryan, M. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Sci Eng Ethics* 26, 2749–2767 (2020).

Stilgoe, J., R. Owen and P. Macnagthen (2012), 'Developing a framework for responsible innovation', *Research Policy*, 42 (9), 1568–80.

Tallacchini, M. (2009). Governing by Values. EU Ethics: Soft Tool, Hard Effects. *Minerva*, *47*(3), 281–306.

Tavani, H. T. (2007) Ethics & Technology. Ethical Issues in an Age of Information Technology (Hoboken: John Wiley & Sons).

Trilateral Research (2022) A survey of artificial intelligence risk assessment methodologies . The global state of play and leading practices identified. Accessed on June 13th at: https://www.trilateralresearch.com/wp-content/uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf

Vanderelst, D., and Winfield, A. (2018) The Dark Side of Ethical Robots. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics and Society (AIES '19). Association for Computing Machinery, New York, 317-322.

Van Dijk, N., Gellert, R., Rommetveit, K. (2016) "A Risk to a Right? Beyond Data Protection Risk Assessments". *Computer Law Security Review*. 32(2), 286-307.

Van Dijk, N. (2021) Tracing Networked Infrastructures of Post-Truth: public dissections of and by techno-political Leviathans. In: Rommetveit, K. (ed., 2021) *Post-Truth Imaginations: New starting points for critique of Politics and Technoscience*. Routledge.

Verbeek, P.-P. (2006). Materializing Morality: Design Ethics and Technological Mediation. *Science, Technology, & Human Values*, *31*(3), 361–380.

Veruggio, G. (2006) The EURON Roboethics Roadmap. Accessed on May 25th 2023 at: http://www.roboethics.org/atelier2006/docs/ROBOETHICS%20ROADMAP%20Rel2.1.1.pdf

Von Schomberg, R. (2011) "Governance and Ethics of Emerging ICT and Security Technologies". Publication series Governance and Ethics, DG Research, European Commission.

Von Schomberg, R. (2013), 'A vision of responsible research and innovation', in R. Owen, J. Bessant, M. Heintz (eds), Responsible Innovation. Managing the Responsible Emergence of Science and Innovation in Society, Chichester: John Wiley & Sons, pp. 51–75.

Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books. 2019.