Future Ethics: Robots, Algorithms and Biases

Anton van Niekerk

Director: Centre for Applied Ethics

Stellenbosch University

Introduction

- The question to be addressed in this paper is: do the technological developments of the 4IR justify – let alone compel – us to acknowledge new forms of ethics or a radically rethinking of the nature of what we currently regard as ethics?
- Probably too early to know
- AI challenges us to rethink our idea/understanding of moral actors
- In the end, it challenges us to retjink our understanding of humanity itself

Context of Debate: 4IR

- What is 4IR?
- “The First Industrial Revolution used water and steam to mechanize production. The Second used electric power to create mass production. The Third used electronics and information technology to automate mass production” (Shwab 2016).
- To this Plutschinski adds: “Now based on a completely digitalized world the 4th Industrial Revolution is characterized by a fusion of technologies that is blurring the lines between the physical, digital and biological spheres. The possibilities will be multiplied by emerging technology breakthroughs in fields such as artificial intelligence, robotics, the Internet of Things, autonomous vehicles, 3-D printing, nanotechnology, biotechnology, material science energy storage and quantum computing” (Plutschinski 2017).

Morality and Ethics: Definitions

- *Morality*, in my understanding, refers to the universal, demonstrable and observable social phenomenon that people of all known cultures submit their behaviour to the demands of obligation. Put more simply: all people that we are aware of, acknowledge and accept that it is a legitimate question to ask whether an action is right or wrong/good or bad.
- E.g.: Universities world-wide denounce plagiarism as phenomenon (i.e. morality; something we observe to be the case.))

- *Ethics*, on the other hand, is the outcome of a more intellectual enterprise, viz. *reflection on the nature of the difference between right and wrong*, as well as the development of argumentative strategies ("theories") in terms of which the difference between right and wrong/good and bad actions can be established and motivated.
- E.g.: Why is it wrong to plagiarise? It compromises the integrity of science. Integrity is an overwhelmingly important moral and scientific value. (Outcome of *ethics as argument*; yields *theory*)

Ethical issue in 4IR to be discussed: **Algorithmic Biases**

- What is an "algorithm"?
- "An algorithm is a methodical set of steps that can be used to make calculations, resolve problems and reach decisions. An algorithm is not a particular calculation, but the method followed when making the calculation (Harari)"
- We use an algorithm to calculate the average of two numbers. Carefully following a cooking recipe to bake a cake, is applying/using an algorithm. Beverage vending machines operate via the application of an algorithm.
- The algorithms controlling vending machines work through mechanical gears and electric circuits. The algorithms controlling humans work through sensations, emotions and thoughts" (Harari 2016, pp. 97-99).

Algorithmic Biases

- Algorithms do not operate neutrally. They reflect the biases, preferences, patterns of the people who develop and manufacture them.

- We are all familiar with how algorithms affect and influence us. A new subscriber to Netflix discovers, within a week or three, that a certain kind of film or program is consistently advertised to him/her, based entirely on an algorithm that has been construed with reference to material that he/she has already watched. This also goes for advertisements on Google of Facebook.

The remarkable, yet simultaneously morally disturbing aspects of the workings of algorithms that are the outcome of AI is, as far as I am concerned, not so much the fact that information is processed on the basis of biases. The inevitability of biases when interpreting hermeneutical objects like texts have been comprehensively argued by authoritative hermeneutical philosophers such as the German Hans-Georg Gadamer (Gadamer 1975: 235-267). For Gadamer, we necessarily interpret symbolic forms on the basis of biases/preconceptions that we already hold.

The significant difference between algorithmic biases and hermeneutic philosophy therefore is not that biases are operational in both, but that the biases cultivated in AI come to the fore purely on the basis of robotic interventions that are operationalised within a few days or weeks, and not, as with Gadamer, on the basis of a lifetime of experiences and pre-judgements. Also, it seems to me that the content of the biases created by AI can be technically manipulated, over and against the biases of the hermeneutic experience that could take years and decades to develop.

Comparison: Human & Robotic experiences

- **<u>Duyndam's questions about</u>**:
- Judgement?
- Hesitation?
- Self-reflection?
- Choices/decisions?
- Mistakes?
- Pardon/absolution?
- Feasting?
- Addiction?
- Restoration of balance/harmony?
- Sympathy & Empathy?
- Future Orientation?
- Again: Judgement – now "Phronesis"!

- Note how each of the factors in terms of which I questioned the compatibility of humanity and robotic behaviour, **link with key dimensions of our understanding of ethics**. To hesitate, to make mistakes, to experience guilt or melancholia, to judge in formalised and non-formalised settings – all of these are dispositions, actions and orientations that are uniquely tied to our ethical consciousness. It is not at all self-evident if and how such dispositions can be translated into the algorithms that drive the behaviour of machine technologies. The human dimension of ethics is not to be underestimated.
- **What, then, is in stall for "a future ethics**" if the developments alluded to earlier actually come to fruition? Firstly, ethics as the outcome of reflection about the difference between right and wrong, will become more important than ever.
- Knowledge, science and technology are not value-free, and this will not change in the future. What will be unusual, is the emergence and prevalence of life-forms or machines imitating life forms and that are capable of harmful behaviour if not well controlled. The control exerted over these phenomena in order to prevent harm will become an ever-increasing part of the ethics of the future.

- What will or ought not to change, is the intuition that the most important category of ethics is *responsibility*. The 4IR is essentially a reflection of the growth of our power over nature and society. More power must mean the acceptance of a more developed and a more focused sense of responsibility.

- One would also hope that the manifest extension of the spheres of influence that the 4IR might open up, will generate a more open and pronounced public discourse and debate about the direction we wish to steer with the help of these new developments. This in particular also pertains to the way in which the benefits of the 4IR will be distributed in society amongst all who can benefit from them.