Ethics and Information Technology (2018) 20:1–3
https://doi.org/10.1007/s10676-018-9450-z

EDITORIAL

Ethics in artificial intelligence: introduction to the special issue

Virginia Dignum[1]

Published online: 13 February 2018

Recent developments in Artificial Intelligence (AI) have generated a steep interest from media and general public. As AI systems (e.g. robots, chatbots, avatars and other intelligent agents) are moving from being perceived as a tool to being perceived as autonomous agents and team-mates, an important focus of research and development is understanding the ethical impact of these systems. What does it mean for an AI system to make a decision? What are the moral, societal and legal consequences of their actions and decisions? Can an AI system be held accountable for its actions? How can these systems be controlled once their learning capabilities bring them into states that are possibly only remotely linked to their initial, designed, setup? Should such autonomous innovation in commercial systems even be allowed, and how should use and development be regulated? These and many other related questions are currently the focus of much attention. The way society and our systems will be able to deal with these questions will for a large part determine our level of trust, and ultimately, the impact of AI in society, and the existence of AI.

Contrary to the frightening images of a dystopic future in media and popular fiction, where AI systems dominate the world and is mostly concerned with warfare, AI is already changing our daily lives mostly in ways that improve human health, safety, and productivity (Stone et al. 2016). This is the case in domain such as transportation; service robots; health-care; education; public safety and security; and entertainment. Nevertheless, and in order to ensure that those dystopic futures do not become reality, these systems must be introduced in ways that build trust and understanding, and respect human and civil rights. The need for ethical considerations in the development of intelligent interactive systems is becoming one of the main influential areas of research in the last few years, and has led to several initiatives both from researchers as from practitioners, including the IEEE initiative on Ethics of Autonomous Systems[1], the Foundation for Responsible Robotics[2], and the Partnership on AI[3] amongst several others.

As the capabilities for autonomous decision making grow, perhaps the most important issue to consider is the need to rethink responsibility (Dignum 2017). Whatever their level of autonomy and social awareness and their ability to learn, AI systems are artefacts, constructed by people to fulfil some goals. Theories, methods, algorithms are needed to integrate societal, legal and moral values into technological developments in AI, at all stages of development (analysis, design, construction, deployment and evaluation). These frameworks must deal both with the autonomic reasoning of the machine about such issues that we consider to have ethical impact, but most importantly, we need frameworks to guide design choices, to regulate the reaches of AI systems, to ensure proper data stewardship, and to help individuals determine their own involvement.

Values are dependent on the socio-cultural context (Turiel 2002), and are often only implicit in deliberation processes, which means that methodologies are needed to elicit the values held by all the stakeholders, and to make these explicit can lead to better understanding and trust on artificial autonomous systems. That is, AI reasoning should be able to take into account societal values, moral and ethical considerations; weigh the respective priorities of values held by different stakeholders in various multicultural contexts; explain its reasoning; and guarantee transparency. Responsible Artificial Intelligence is about human responsibility for the development of intelligent systems along fundamental human principles and values, to ensure human flourishing and wellbeing in a sustainable world. In fact, Responsible AI is more than the ticking of some ethical 'boxes' in a report, or the development of some add-on features, or switch-off buttons in AI systems. Rather, responsibility is fundamental

✉ Virginia Dignum
m.v.dignum@tudelft.nl

[1] Delft Design for Values Institute, Delft University of Technology, Jaffalaan 5, 2628BX Delft, The Netherlands

[1] http://ethicsinaction.ieee.org
[2] http://responsiblerobotics.org/
[3] http://www.partnershiponai.org/

to autonomy and should be one of the core stances underlying AI research.

The above considerations show that ethics and AI are related at several levels:

- Ethics *by* Design: the technical/algorithmic integration of ethical reasoning capabilities as part of the behaviour of artificial autonomous system;
- Ethics *in* Design: the regulatory and engineering methods that support the analysis and evaluation of the ethical implications of AI systems as these integrate or replace traditional social structures;
- Ethics *for* Design: the codes of conduct, standards and certification processes that ensure the integrity of developers and users as they research, design, construct, employ and manage artificial intelligent systems.

The papers on this special issue present different views on the relation between ethics and AI. The first two papers, those by Rahwan, and by Bryson, can be classified mostly in the area of Ethics *in* Design and for parts in the area of Ethics *for* Designers, whereas the last three papers, by Vamplew et al., Bonnemains et al., and Arnold and Scheutz propose different approaches to Ethics *by* Design.

The paper by Iyad Rahwan, "Society-in-the-Loop: Programming the Algorithmic Social Contract", focuses on the regulatory and governance mechanisms for autonomous machines. The vision of the paper is that the algorithms governing our lives must be provably transparent, fair, and accountable along the values shared by stakeholders. The paper describes a conceptual framework to program, debug and maintain an algorithmic social contract, a pact between various human stakeholders, mediated by machines, herewith setting the *society-in-the-loop* (SITL) approach for identifying and negotiating the values of various stakeholders affected by AI systems, as basis for monitoring compliance of the system with the social contract.

In "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics", Joanna Bryson contends that the place of AI in society is a matter of normative, rather than descriptive ethics. In the view exposed in this paper, the question of whether AI or robots can, or should, be afforded moral agency or patiency is not one amenable either to discovery or simple reasoning, because we as societies constantly reconstruct our artefacts, including our ethical systems. Taking a functionalist assumption, that ethics is the set of behaviour that maintains a society, the paper explores the basis of sociality and autonomy to explain moral intuitions with respect to AI systems. This effort leads to the conclusion that while constructing AI as either moral agent or patient is possible, neither is desirable, given the unlikelihood of constructing a suitable coherent ethics of AI moral subjectivity. The paper presents solid arguments to Bryson's position ther "We are therefore obliged not to build AI we are obliged to".

The second set of papers, focus on the issue of Ethics by Design. I.e. assuming that designers are given a clear, consistent and share set of ethical principles, these three papers propose different aspects of its implementation in AI systems, such that the system is able either to make ethically-based decisions itself, or to alert users and/or monitors to potential deviations of behaviour from such ethical principles.

Peter Vamplew et al. focus on the need to ensure that the behaviour of AI systems is beneficial to humanity. In their paper "Human-Aligned Artificial Intelligence is a Multiobjective Problem", they discuss the requirement for ethical, legal and safety-based frameworks to consider multiple potentially conflicting factors. They demonstrate that these alignment frameworks can be represented as utility functions, but that the widely used Maximum Expected Utility (MEU) paradigm provides insufficient support for such multiobjective decision-making. They then propose a Multiobjective Maximum Expected Utility paradigm based on the combination of vector utilities and non-linear action-selection that can overcome many of the issues which limit MEU's effectiveness in implementing values-aligned artificial intelligence. They further examine existing approaches to multiobjective artificial intelligence, and identify how these can contribute to the development of human-aligned intelligent agents.

In "Embedded Ethics: Some technical and ethical challenges", Vincent Bonnemains, Claire Saurel and Catherine Tessier focus on a formal approach to what can be considered as artificial ethical reasoning by an observer. The approach includes formal tools to describe a situation and models of ethical principles that are designed to automatically compute a judgement, and to explain why a given decision is ethically, or not, acceptable. Based on a though experiment involving the drone dilemma, the paper illustrates the use of this approach to model three ethical frameworks—utilitarian ethics, deontological ethics and the Doctrine of Double effect—and evaluate their responses to this ethical dilemma.

Finally, the paper "The Big Red Button Is Too Late: An Alternative Model for the Ethical Evaluation of AI Systems", by Thomas Arnold and Matthias Scheutz presents existing proposals for an *emergency button* in AI systems, and discuss the viability of emergency stop mechanisms that enable human operators to interrupt or divert a system while preventing the system from learning that such an intervention is a threat to its own existence. Given that such approaches concentrate on minimizing effects after the system has already gone astray, the paper proposes an alternative based on an ongoing self-evaluation and testing an integral part of a system's operation, to prevent chaos and risk

before they start and diagnose how the system is in error. The paper further argues for a scenario-generation mechanism that enables to test a system's decisions in a simulated world, rather than the real world, which they conclude to be far more effective, responsive, and vigilant toward a system's learning and action in the world than an emergency button which one might not get to push in time.

Together, these papers represent current state of the art in Ethics in Artificial Intelligence, and contribute to a better understanding of the many challenges for this topic.

References

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'2017)*, pp. 4698–4704

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., Teller, A. (2016). Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel.

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. Cambridge: Cambridge University Press.