PERSPECTIVES

ARTIFICIAL INTELLIGENCE

How AI can be a force for good

An ethical framework will help to harness the potential of AI while keeping humans in control

By **Mariarosaria Taddeo**[1,2,3]
and **Luciano Floridi**[1,2]

Artificial intelligence (AI) is not just a new technology that requires regulation. It is a powerful force that is reshaping daily practices, personal and professional interactions, and environments. For the well-being of humanity it is crucial that this power is used as a force of good. Ethics plays a key role in this process by ensuring that regulations of AI harness its potential while mitigating its risks.

AI may be defined in many ways. Get its definition wrong, and any assessment of the ethical challenges of AI becomes science fiction at best or an irresponsible distraction at worst, as in the case of the singularity debate. A scientifically sound approach is to draw on its classic definition (*1*) as a growing resource of interactive, autonomous, self-learning agency, which enables computational artifacts to perform tasks that otherwise would require human intelligence to be executed successfully (*2*). AI can then be further defined in terms of features such as the computational models on which it relies or the architecture of the technology. But when it comes to ethical and policy-related issues, the latter distinctions are unnecessary (*3*). On the one hand, AI is fueled by data and therefore faces ethical challenges related to data governance, including consent, ownership, and privacy. These data-related challenges may be exacerbated by AI, but would occur even without AI. On the other hand, AI is a distinct form of autonomous and self-learning agency and thus raises unique ethical challenges. The latter are the focus of this article.

TOMORROW'S EARTH
Read more articles online at scim.ag/TomorrowsEarth

The ethical debate on AI as a new form of agency dates to the 1960s (*2*, *4*). Since then, many of the relevant problems have concerned delegation and responsibility. As AI is used in ever more contexts, from recruitment to health care, understanding which tasks and decisions to entrust (delegate) to AI and how to ascribe responsibility for its performance are pressing ethical problems. At the same time, as AI becomes invisibly ubiquitous, new ethical challenges emerge. The protection of human self-determination is one of the most relevant and must be addressed urgently. The application of AI to profile users for targeted advertising, as in the case of online service providers, and in political campaigns, as unveiled by the Cambridge Analytica case, offer clear examples of the potential of AI to capture users' preferences and characteristics and hence shape their goals and nudge their behavior to an extent that may undermine their self-determination.

DELEGATION AND RESPONSIBILITY

AI applications are becoming pervasive. Users rely on them to deal with a variety of tasks, from delivering goods to ensuring national defense (*5*). Assigning these tasks to AI brings huge benefits to societies (see the photo). It lowers costs, reduces risks, increases consistency and reliability, and enables new solutions to complex problems. For example, AI applications can lower diagnostic errors by 85% in breast cancer patients (*6*), and AI cybersecurity systems can reduce the average time to identify and neutralize cyberattacks from 101 days to a few hours (*5*).

However, delegation may also lead to harmful, unintended consequences, especially when it involves sensitive decisions or tasks (*7*, *8*) and excludes or even precludes human supervision (*3*). The case of COMPAS, an AI legal system that discriminated against African-American and Hispanic men when making decisions about granting parole (*9*), has become infamous. Robust procedures for human oversight are needed to minimize such unintended consequences and redress any unfair impacts of AI.

Still, human oversight is insufficient if it deals with problems only after they occur. Techniques to explain AI and predict its outcomes are also needed. The Explainable Artificial Intelligence program of DARPA (Defense Advanced Research Project Agency) is an excellent example. The goal of this program is to define new techniques to explain the decision-making processes of AI systems. This will enable users to understand how AI systems work, and designers and developers to improve the systems to avoid mistakes and mitigate the risks of misuse. To be successful, similar projects must include an ethical impact analysis from the beginning, to assess AI's benefits and risks and define guiding principles for an ethically sound design and use of AI.

The effects of decisions or actions based on AI are often the result of countless interactions among many actors, including designers, developers, users, software, and hardware. This is known as distributed agency (*10*). With distributed agency comes distributed responsibility. Existing ethical frameworks address individual, human responsibility, with the goal of allocating punishment or reward based on the actions and intentions of an individual. They were not developed to deal with distributed responsibility.

Only recently have new ethical theories been defined to take distributed agency into account. The proposed theories rely on contractual and tort liability (*11*) or on strict liability (*12*) and adopt a faultless responsibility model. This model separates responsibility of an agent from their intentions to perform a given action or their ability to control its outcomes, and holds all agents of a distributed system, such as a company, responsible. This is key when considering the case of AI, because it distributes moral responsibility among designers, regulators, and users. In doing so, the model plays a central role in preventing evil and fostering good, because it nudges all involved agents to adopt responsible behaviors.

Establishing good practices for delegation and defining new models to ascribe moral responsibility are essential to seize the opportunities created by AI and address the related challenges, but they are still not enough. Ethical analyses must be extended to account for the invisible influence exercised by AI on human behavior.

INVISIBILITY AND INFLUENCE

AI supports services, platforms, and devices that are ubiquitous and used on a daily basis. In 2017, the International Federation of Robotics suggested that by 2020, more than 1.7 million new AI-powered robots will be installed in factories worldwide. In the same year, the company Juniper Networks issued a report estimating that, by 2022, 55% of households worldwide will have a voice assistant, like Amazon Alexa.

As it matures and disseminates, AI blends into our lives, experiences, and environ-

[1]*Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK.* [2]*The Alan Turing Institute, 96 Euston Road, London NW1 2DB, UK.* [3]*Department of Computer Science, University of Oxford, Oxford OX1 3QD, UK. Email: mariarosaria.taddeo@oii.ox.ac.uk*

ments and becomes an invisible facilitator that mediates our interactions in a convenient, barely noticeable way. While creating new opportunities, this invisible integration of AI into our environments poses further ethical issues. Some are domain-dependent. For example, trust and transparency are crucial when embedding AI solutions in homes, schools, or hospitals, whereas equality, fairness, and the protection of creativity and rights of employees are essential in the integration of AI in the workplace. But the integration of AI also poses another fundamental risk: the erosion of human self-determination due to the invisibility and influencing power of AI.

This invisibility enhances the influencing power of AI. With their predictive capabilities and relentless nudging, ubiquitous but imperceptible, AI systems can shape our choices and actions easily and quietly. This is not necessarily detrimental. For example, it may foster social interaction and cooperation (*13*). However, AI may also exert its influencing power beyond our wishes or understanding, undermining our control on the environment, societies, and ultimately on our choices, projects, identities, and lives. The improper design and use of invisible AI may threaten our fragile, and yet constitutive, ability to determine our own lives and identities and keep our choices open.

The Avatar Kids project allows hospitalized children to be present in the classroom through a remote-controlled robot.

TRANSLATIONAL ETHICS

To deal with the risks posed by AI, it is imperative to identify the right set of fundamental ethical principles to inform the design, regulation, and use of AI and leverage it to benefit as well as respect individuals and societies. It is not an easy task, as ethical principles may vary depending on cultural contexts and the domain of analysis. This is a problem that the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (*14*) tackles with the aim of advancing public debate on the values and principles that should underpin ethical uses of AI.

More important, some agreement on the fundamental principles is emerging. A recent comparative analysis (*15*) of the main international initiatives focusing on AI ethics highlights substantive overlap of the principles endorsed by these initiatives and some of the key principles of bioethics, namely beneficence, nonmaleficence, autonomy, and justice. There is reason to be optimistic about further convergence, as other principles may be extracted from the Universal Declaration of Human Rights. This convergence will foster coherence, and hence compatibility, of different ethical frameworks for AI and provide overarching ethical guidance for the design, regulations, and uses of this technology.

Once identified, ethical principles must be translated into viable guidelines to shape AI-based innovation. Such translation has precedents, especially in medicine, where translational research goes "from bench to bedside," building on research advances in biology to develop new therapies and treatments. Likewise, translational ethics builds on academic advances to shape regulatory and governance approaches. This approach underpins the forthcoming recommendations for the ethical design and regulation of AI to be issued by the AI4People project.

Launched in the European Parliament in February 2018, AI4People was set up to help orient AI toward the good of society and everyone in it. The initiative combines efforts of a scientific committee of international experts and a forum of stakeholders, in consultation with the High-Level Expert Group on Artificial Intelligence of the European Commission, to propose a series of concrete and actionable recommendations for the ethical and socially preferable development of AI.

A translational ethics of AI needs to formulate foresight methodologies to indicate ethical risks and opportunities and prevent unwanted consequences. Impact assessment analyses are an example of this methodology. They provide a step-by-step evaluation of the impact of practices or technologies deployed in a given organization on aspects such as privacy, transparency, or liability.

Foresight methodologies can never map the entire spectrum of opportunities, risks, and unintended consequences of AI systems, but may identify preferable alternatives, valuable courses of action, likely risks, and mitigating strategies. This has a dual advantage. As an opportunity strategy, foresight methodologies can help leverage ethical solutions. As a form of risk management, they can help prevent or mitigate costly mistakes, by avoiding decisions or actions that are ethically unacceptable. This will lower the opportunity costs of choices not made or options not seized for lack of clarity or fear of backlash.

Ethical regulation of the design and use of AI is a complex but necessary task. The alternative may lead to devaluation of individual rights and social values, rejection of AI-based innovation, and ultimately a missed opportunity to use AI to improve individual well-being and social welfare. Humanity learned this lesson the hard way when it did not regulate the impact of the industrial revolution on labor forces, and also when it recognized too late the environmental impact of massive industrialization and global consumerism. It has taken a very long time, social unrest, and even revolutions to protect workers' rights and establish sustainability frameworks.

The AI revolution is equally significant, and humanity must not make the same mistake again. It is imperative to address new questions about the nature of post-AI societies and the values that should underpin the design, regulation, and use of AI in these societies. This is why initiatives like the abovementioned AI4People and IEEE projects, the European Union (EU) strategy for AI, the EU Declaration of Cooperation on Artificial Intelligence, and the Partnership on Artificial Intelligence to Benefit People and Society are so important (see the supplementary materials for suggested further reading). A coordinated effort by civil society, politics, business, and academia will help to identify and pursue the best strategies to make AI a force for good and unlock its potential to foster human flourishing while respecting human dignity. ■

REFERENCES AND NOTES

1. J. McCarthy *et al.*, *AI Mag.* **27**, 12 (2006).
2. A. L. Samuel, *Science* **132**, 741 (1960).
3. G.-Z. Yang *et al.*, *Sci. Robot.* **3**, eaar7650 (2018).
4. N. Wiener, *Science* **131**, 1355 (1960).
5. M. Taddeo, L. Floridi, *Nature* **556**, 296 (2018).
6. D. Wang *et al.*, arXiv:1606.05718 [q-bio.QM] (18 June 2016).
7. P. Asaro, *Int. Rev. Red Cross* **94**, 687 (2012).
8. S. Russell, *Nature* **521**, 415 (2015).
9. J. Larson *et al.*, "How We Analyzed the COMPAS Recidivism Algorithm" (May 2016); www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
10. L. Floridi, *Sci. Eng. Ethics* **19**, 727 (2013).
11. U. Pagallo, in *Human Law and Computer Law: Comparative Perspectives*, M. Hildebrandt, J. Gaakeer, Eds. (Springer, Netherlands, 2013), pp. 47–65.
12. L. Floridi, *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* **374**, 20160112 (2016).
13. H. Shirado, N. A. Christakis, *Nature* **545**, 370 (2017).
14. IEEE Standards Association, Ethically Aligned Design, Version 2.
15. J. Cowls, L. Floridi, Prolegomena to a White Paper on an Ethical Framework for a Good AI Society, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3198732, 19 June 2018.

ACKNOWLEDGMENTS

M.T. and L.F. are members of the Partnership on Artificial Intelligence to Benefit People and Society; L.F. is also chair of the scientific committee of AI4People.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/361/6404/751/suppl/DC1

10.1126/science.aat5991