Final Statement on Robotics, AI, and Humanity: Science, Ethics, and Policy

A concluding statement from the Conference on "Robotics, AI and Humanity, Science, Ethics and Policy" organized jointly by the Pontifical Academy of Sciences (PAS) and the Pontifical Academy of Social Sciences (PASS)

Issues and Agenda

1. Advances in machine learning (often referred to as artificial intelligence – AI) and robotics are accelerating. They significantly impact upon the functioning of societies and economies, and have prompted widespread debate over their benefits and drawbacks for humanity. This fast-moving field of science and technology requires our careful attention. The Pontifical Academies have accordingly organized this conference,1 which brought together colleagues from different disciplines.

2. Artificial intelligence and applications in robot technologies will have far-reaching implications for economies, the fabric of society, and culture. In addition to examining the current research frontiers in AI/robotics, we reviewed and discussed the likely impacts on societal well-being, and ethical and policy implications.

3. AI and robotics hold great promise to address some of our most intractable social, economic and environmental problems, such as climate change and achieving the UN's Sustainable Development Goals (SDGs) for 2030.

4. There are also long-term trends in AI and robotics, with consequences that may ultimately challenge the place of humans in society.

5. Of growing concern are the risks for peace due to new forms of warfare (cyber-attacks, autonomous weapons), calling for new international security regulations.

6. Ethical and religious aspects of AI and robotics need clarification in order to guide potential needs for regulatory policies on applications and the future development of AI/robotics.

Foundational Issues in AI and Robotics

7. The field of AI has developed a rich variety of **theoretical approaches and frameworks** on the one side, and increasingly impressive **practical applications** on the other. AI has the potential to bring about advances in every area of science and society. It may help us overcome some of our cognitive limitations and solve complex problems. While vast amounts of data present a challenge to human cognitive abilities, Big Data presents unprecedented opportunities for science and the humanities. The translational potential of Big Data is considerable, for instance in medicine, public health, education, and the management of complex systems in general (biosphere, geosphere, economy). However, the science based on big data as such remains empiricist and challenges us to discover the underlying causal mechanisms generating patterns.

8. In combination with **robotics and brain-computer interfaces**, AI already brings unique support to patients with sensory or motor deficits and facilitates caretaking of disabled patients. By providing novel tools for knowledge acquisition, AI may bring dramatic changes in education and facilitate access to knowledge. There may also be synergies arising from robot-to-robot interaction and potential positive synergy of humans and robots working together on tasks.

9. Among the foundational issues of AI and robotics is the question of whether machines may hypothetically attain **capabilities such as consciousness**. This is currently debated from the contrasting perspectives of natural science, social theory, and philosophy; this remains an open issue, in large measure because there is diversity of definitions of “consciousness”. Questions arise whether the emphasis on the AI’s supra-human capacities for computation and compilation masks the many limitations of artificial systems.

10. Most maintain that **robots cannot be considered as persons**, so robots will not and should not be free agents or possess rights. Some, however, argue, ‘command and control’ conceptions may not be appropriate to human-robotic relations, and others even ask if something like ‘electronic citizenship’ may be considered. **Christian philosophy and theology** states that the human soul is intrinsically incorruptible. This is the metaphysical foundation according to which the human person is in himself or herself free and capable of ethical order, and emerges from the forces of nature. As a spiritual subject, the human being is imago Dei. In this sense of Christian philosophy, AI/Robots cannot be considered as persons, so robots will not and should not possess human freedom and not possess a spiritual soul and cannot be considered “images of God” but maybe “images of human beings”, and can be created by humans to be their instruments for the good of human society.

11. Within machine learning research, there is a line of development that aims to identify foundational justifications for the **design of cognitive agents**; such justifications would enable the derivation of theorems characterizing the possibilities and limitations of intelligent agents. **Cognitive agents** act within an open, partially or completely unknown environment in order to achieve goals. Key concepts for a foundational framework for AI include: agents, environments, rewards, local scores, global scores, the exact model of interaction between agents and environments, and a specification of the available computational resources of agents and environments. An "intelligent agent" may be defined as an agent that can achieve goals in a wide range of environments. The topic of cognition with bounded resources merits further exploration in order to reach a meaningful and broadly integrated foundational framework for artificial intelligence.

12. A central aspect of learning from experience is the representation and processing of uncertain knowledge. In the absence of deterministic assumptions about the world, there is no nontrivial logical conclusion that can be drawn from the past for any future event. Accordingly, it is of foundational interest to analyze the **structure of uncertainty** as a question in its own right. Much research remains to be conducted in order to translate preliminary proposals and formal methods from the theoretical realm into the engineering of efficient algorithmic solutions.

13. Some recent results establish a tight **connection between learnability and provability**, thus reducing the question of what can be effectively learned to the foundational questions of mathematics with regard to set existence axioms. Results of reverse mathematics, a branch of mathematical logic analyzing theorems with reference to the set existence axioms necessary to prove them may be used to illustrate the implications of machine learning frameworks. In general, model checking and proof checking techniques become ever more important as the criticality of tasks entrusted to intelligent agents is expanded.

14. Until recently, basic mathematical science had few (if any) ethical issues on its agenda. However, given that mathematicians and software designers are central to the development of AI, it is essential that they consider the ethical implications of their work. In light of the questions that are increasingly raised about the trustworthiness of autonomous systems**, AI developers have a responsibility – that ideally should become a legal obligation – to create trustworthy and controllable robot systems**.

The Science, Engineering and AI/Robot-Human Interactions

15. Major research is underway in areas that define us as humans, such as language, symbol processing, one-shot learning, self-evaluation, confidence judgment, program induction, conceiving goals, and integrating existing modules into an overarching, multi-purpose intelligent architecture. Computational agents trained by reinforcement learning and deep learning frameworks demonstrate outstanding performance in tasks heretofore thought intractable. While a thorough foundation for a general theory of computational cognitive agents is still missing, the

conceptual and practical advance of artificial intelligence has reached a state in which **ethical and safety questions and the impact on society overall become pressing issues**. For example, AI-based inferences of persons' feelings derived from face recognition data is such an issue.

16. The spread of robotics profoundly modifies human and social relations in all spheres of society, in the family as well as in the workplace and in the public sphere. These modifications take on the character of hybridization processes between the properly human characteristics of relationships and the artificial ones, hence between analogical and virtual reality. Therefore, it is necessary to increase scientific research on issues concerning the social effects that derive from delegating relevant aspects of the social organization to AI and robots. An aim of this research should be to understand how it is possible to govern the relevant processes of change and produce those relational goods that realize a virtuous human fulfillment within a sustainable and fair societal development

17. We note **fast progress in robotics engineering** is transforming whole industries (industry 4.0). The evolution of the internet of things (IoT) with communication among machines and inter-connected machine learning offers major advances for services such as banking and finance. Robot-robot and human-robot interactions are increasingly intensive. Yet, AI systems are hard to test and validate. This makes trust in AI and robots challenging. **Issues of regulations** and ownership of data, of assignment of responsibilities and transparency of algorithms are arising, and require legitimate institutional arrangements.

18. We can distinguish between mechanical robots, designed to accomplish routine tasks in production, and AI's capacities to assist in social care, medical procedures, safe and energy efficient mobility systems, educational tasks, and scientific research. While intelligent assistants may benefit adults and children alike, they also carry risks because their impact on the developing brain is unknown, and because **people may lose motivation in areas where AI appear superior**.

19. Over the past decades the field of robotics has spurred **a multitude of applications of novel services and assistance.** Paradigmatic for many application scenarios are robotic hand-arm systems for which the challenges of precision, sensitivity and robustness come along with safe grasping requirements. Although robotic hands are still a long way from their human counterparts their performance has been greatly enhanced by new control methods. Promising applications are evolving in tele-manipulation systems in a variety of areas such as healthcare, factory production, and mobility.

20. **AI may serve good governance**, including the identification, and prevention of illegal transactions, for instance money received from criminal activities such as drug trafficking, human trafficking or illegal transplants. However, when AI is in the hands of companies alone, the revenues from AI may not be redistributed equitably. These new technologies must not become instruments to enslave people or further marginalize the poor.

Robotics Changing the Future of Work, Farming, Poverty, Inequality and Ecology

21. We reviewed AI (and related emergent technologies) applications in medicine and health care, mobility and transport, manufacturing, and agriculture. Major opportunities were noted, and considerable attention was devoted to robotics/AI applications in each of these domains considered separately. However, a sectorial perspective on AI and robotics has limitations. It seems necessary to gain a fuller picture of the **connections between the applications** and a focus on public policies that facilitate overall **fairness and equity** covering all aspects of AI and Robotics.

22. Unless channeled for public benefit, AI may raise important concerns for the economy and the stability of society. **Jobs may be lost** to computerized devices, with a resulting increase in income disparity and knowledge gaps. Advances in automation and increased supplies of artificial labor particularly in the agricultural and industrial sectors can significantly reduce employment in emerging economies. Through linkages within global value chains, workers in low-income countries may also be affected by robots in higher-income countries, which could reduce the need for outsourcing routine jobs to the former low-wage regions. However, robot use could also increase the demand for labor by reducing the cost of production which leads to industrial expansion. Reliable estimates of new jobs created in the industries designing and manufacturing robots are lacking but are needed.

23. The employment and work implications of robotics is a major public policy issue. Policies should aim at **providing the necessary social security measures** for affected workers while investing in the development of the necessary skills to take advantage of the new jobs created. The state should be able to redistribute the profits that are earned from the work carried out by robotics. Such redistribution could, for instance, pay for the re-training of affected individuals so that they can remain within the work force. In this connection, it is important to remember that many of these new technological innovations have been achieved with support from public funding.

24. **Robots, AI, and digital capital in general can be considered as a tax base**. Currently this is not the case; human labor is taxed but robotic labor is not. In this way, robotic systems are indirectly subsidized, as companies can offset them in their accounting systems, thus reducing corporate taxation. These distortions that dis-favor human workers, while favoring investment in robots, should be reversed.

25. We note that implications, opportunities and risks of **AI and Robotics for sustainable development and the poor need more attention.** Especially implications for low -income countries, the marginalized, and women need study and consideration in programs and policies. AI teaching resources in many low income regions are an opportunity. As a large proportion of the poor live on small farms, particularly in Africa and South and East Asia, it matters whether or not they get access to meaningful digital technologies and AI. Examples are land ownership certification through blockchain technology; precision technologies in land and crop management and many

more.

26. Direct and indirect **environmental impacts** should be considered more. Monitoring through smart remote sensing in terrestrial and aquatic systems can be much enhanced to assess biodiversity change and impacts of interventions. However, there is also the issue of pollution by electronic waste dumped by industrialized countries in low income countries. This needs urgent attention as does the carbon foot print of AI and robotics.

Robotics, AI, and Militarized Conflict

27. Within militarized conflict, AI-based systems (including robots) can serve a variety of purposes, inter alia, extracting wounded personnel, monitoring compliance with laws of war/rules of engagement, improving situational awareness/battlefield planning, and making targeting decisions. While it is the last category that raises the most **challenging moral issues**, in all cases the implications of potentially lowered barriers of war as well as systems risks must be carefully examined before any implementation in battlefield settings.

28. Worries about falling behind in the race to develop new military applications must not become an excuse for short-circuiting safety research, testing, and adequate training. Because weapon design is trending away from large scale infrastructure toward autonomous, de-centralized, and miniaturized systems, risks of destructive effects due to systems designs and potential failure will be greatly magnified in relation to most systems operative today. This increased potential for negative externalities must be compensated by proportionate investment in safety and training. AI tools should enhance but not detract from the exercise of sound and moral judgment by military personnel.

29. Concerted **international effort** must be directed toward identifying the specific applications of AI that pose risks of escalation. States should agree on concrete steps to reduce the risk of AI-facilitated and possibly escalated wars. Attention should not only be paid to the dangers of technology replacing people in military spheres, but also to the danger that AI may pose for the exercise of “strategic reflection” in conflict settings.

30. With respect to lethal autonomous weapons systems, given the present state of technical competence (and for the foreseeable future), no systems should be deployed that function in an unsupervised mode. Lines of **human accountability must be maintained** so that adherence to internationally recognized laws of war can be assured and violations sanctioned.

Society, Ethical, Religious, and Regulatory Dimensions of Robotics/AI

In addition to those already highlighted, the following issues should be emphasized:

31. The efforts of publicly supported development of **intelligent machines should be directed to the common good. The impact on public goods and services**, as well as health, education, happiness and sustainability, must be paramount. AI may have unexpected biases or inhuman consequences

including segmentation of society and racial and gender bias, and these need to be addressed before they might occur. These are national and global issues and the latter need further attention from the United Nations.

32. Regarding privacy, access to new knowledge, and information rights**, the poor are particularly threatened** because of their current lack of power and voice. AI and Robotics needs to be accompanied by more empowerment of the poor through information and education and investment in skills needing enhancement.

33. **The issue of work** is at the centre of the Social Doctrine of the Church. As stated in *Laborem exercens*, access to meaningful work is fundamental to human dignity. Eliminating such access by the use of machines is not acceptable, whilst reducing work burdens and the health risks of work represents a constructive development.

34. Policies should aim for **sharing the benefits of productivity** growth through a combination of profit sharing, not by subsidizing robots but through considering (digital) capital taxation, and a reduction in working time spent on routine tasks.

35. Risks of **manipulative applications** of AI for shaping public opinion and electoral interference need attention and national and international controls are called for.

36. AI and robotics offer great opportunities and entail risks; therefore **regulations should be appropriately designed** by legitimate public institutions, not hampering opportunities and not stimulating excessive risk taking and bias. This needs a framework in which inclusive public societal discourse is informed by sciences of different disciplines, and all segments of society are able to participate.

37. New forms of regulating the digital economy are called for that **ensure proper data protection and personal privacy**. Moreover, deontic values such as ‘permitted’, ‘obligatory’, ‘forbidden’ need to be strengthened to navigate the web and interact with robots.

38. **Companies should create ethical and safety boards**, and join with non-profit organizations that aim to establish best practices and standards for the beneficial deployment of AI and Robotics. Appropriate protocols for AI robots’ safety need to be developed, such as duplicated checking by independent design teams of an AI robot systems. The passing of ethical and safety tests, evaluating for instance the social impact or covert racial prejudice, should become a prerequisite for the release of new AI software. External civil boards performing recurrent and transparent evaluation of all technologies, including the military, should be considered.

39. **Scientists and engineers**, as the designers of AI and robot devices, bear a responsibility in actively trying to ensure that their inventions and innovations are safe and can be used for good

purposes.

[1] A previous conference (30 November–1 December 2016) was organized by the two Academies on the Power and Limits of Artificial Intelligence. The present statement builds on the work of this earlier conference (including the statement that was issued at its conclusion).

(Statement prepared by Joachim von Braun (PAS President) with significant inputs by the Conference Participants).