Ethical Implications of Self-Improving Intelligent Robots

Applicants

Eligible proposals must have two (and only two) applicants from different disciplines within the Network Institute.

Supervisor Name	Department/Group	Faculty
1. Jacqueline Heinerman	Computer Science /Computational Intelligence	FEW
2. Scott Robbins	Theoretical Philosophy/Science Beyond Scientism	FGW

Project description

Provide a brief description of the project (max. 300 words)

There is substantial concern in the general public and even some experts about the perceived threat of artificial intelligence (AI) running amok, improving itself to outstrip human intelligence and negating any possibility of shutting it down. This proposal sets out a research project to investigate and test the safety and controllability of such 'smart' AI, in particular in intelligent robots.

Task specific ('narrow') AI has resulted in systems that outperform humans on specific tasks (e.g., IBM's Deep Blue and Watson programmes). It is debatable whether this actually constitutes true intelligence or if its success is due to sheer computing power rather than to understanding. This debate led to research into 'general' AI that can generalise its skills to new tasks, incorporating mechanisms for learning, generalisation and self-improvement. Self-improving AI can ultimately lead to robots that decide autonomously, on the basis of their own (not necessarily pre-programmed) motivations on their course of action.

This project aims to investigate the technical possibilities and philosophical implications of ensuring that general AI is safe and controllable by achieving the following objectives:

- To draft an inventory of safety, verification and control mechanisms in the context of self-improving general (robotic) AI;
- To validate relevant techniques in the context of self-improving robots in the DREAM project[1];
- Articulate the ethical implications of self-learning, autonomous robots who create their own desires and decide for themselves what they are going to do;
- Propose limits, based on ethical considerations, which should be embedded into the robots to ensure that they do not perform immoral actions;

[1] DREAM (http://robotsthatdream.eu/) is an EU-funded Horizon 2020 project that aims to incorporate sleep and dream-like processes within a robot's cognitive architecture. This enables robots to consolidate their experiences into more useful and generic formats, improving their future ability to learn and adapt.

- Compare the relevant philosophical concepts (i.e. knowledge, learning, desire, motivation, etc.) with the actual implementations in AI systems.

Project Organization

Each proposal requests two Academy Assistants from different disciplines. Describe their roles and describe the skills and expertise required from them. (max. 300 words)

The project requires two students who, besides the usual skills that make a project successful (teamwork, motivation, pro-active attitude), share interest in interdisciplinary research and the ethical implications of general AI.

The Assistant from the Computer Science discipline will (1) investigate current research into self-improving general AI (eg., developmental robotics [Cangelosi2015], knowledge restructuring [Duro2014], etc) and (2) make an inventory of relevant safety, verification and control mechanisms (eg., (semi-)Markov models or probabilistic finite state machines, ethical action selection [Winfield2014]).
(S)he will implement selected mechanisms and experimentally investigate them in robotic experiments.
Good programming skills are essential, but (s)he must also look beyond the technical challenges. Working knowledge of relevant AI and Machine Learning techniques is beneficial.

The assistant from the Philosophy discipline will do a literature study on (1) the contemporary debate on robot ethics (including eg., ethics and robot design [vanWynsberghe2014], robot emotions [Coeckelbergh2010], and robot morality [Wallach2008]), and (2) make an inventory of the relevant theoretical concepts (eg., “knowledge”, “intelligence”, “autonomy”) which the robots are purported to have as well as philosophical interpretations of those concepts which may be compatible with artificial agents. The assistant will use the literature study on robot ethics to propose limits to the autonomous decision making capability of the robots. Differences in the philosophical interpretations of the relevant concepts and the actual implementation of those concepts into the robots will be highlighted.

The assistants will jointly validate the ethical implications and appropriateness of relevant techniques in the context of self-improving robots. They will identify gaps between proposed safety techniques and the requirements posed by ethical considerations in the context of self-improving and generalising AI.

Collaboration

Describe how your research improves collaboration and cross-pollination between the disciplines involved (max. 300 words)

AI research is typically driven by practice, investigating methods to render technology smarter and more capable. While AI researchers are certainly not blind to the ethical and societal impacts of their research, this is usually more of an afterthought and (too) rarely comes to the fore. Philosophy, on the other hand, has a long-standing tradition investigating exactly these impacts in general. Joint research as we propose here will allow philosophers to take a better informed view of the relevance of ethical and societal issues by highlighting how philosophical

concepts are interpreted and implemented by computer scientists. Similarly, the AI field will benefit from the rigorous analysis of the ethical and societal impacts that philosophy affords.

One of the expected outputs is a PhD proposal suitable for submission to agencies such as STW or NWO. This will result, also in the longer term, in joint research that is highly relevant to the field of Robotics and Ethics (aka “Roboethics”). This will position VU researchers so that they can contribute substantially to this rapidly growing field of research.

Deliverables

Enumerate intended project results: papers, research proposals or otherwise. (max 200 words)

The project will yield the following deliverables:

1. An overview of existing methods for safety, verification and control in self-improving general (robotic) AI and their relevance to ethical and societal issues. If there is sufficient body to this research, this may be published as an overview paper;
2. Proof-of-concept implementations of selected methods for safety, verification and control in self-improving general (robotic) AI. If trials with these implementations show promising results, this may result in a publication or be incorporated into the DREAM project’s cognitive architecture;
3. A (joint) PhD proposal. This will lay the basis for cohesive future collaboration in Roboethics research.

Planning

Provide a breakdown of the project into phases with tentative timing (max 150 words)

Task	**Month**	**Responsible**
Literature study into existing initiatives and relevant safety techniques in AI.	1-2	Comp Sci student
Literature study into the ethical implications of self-learning, autonomous robots	1-2	Phil student
Draft overview and select relevant safety techniques and experimental protocol	3-5	both
Develop proof-of-concept implementation	4-6	Comp Sci student
Experimental validation, including a paper write-up if results are interesting enough	5-9	both
PhD proposal write-up	9-10	both

Please respect the word count limits: proposals that exceed the stated limits will not be eligible.

References

[Cangelosi2015] A Cangelosi, M Schlesinger (2015) *Developmental Robotics - From Babies to Robots.* MIT Press.

[Coeckelbergh2010] Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. Ethics and Information Technology, 12(3), 235-241.

[Duro2014] R Duro, JA Becerra, R Salgado, F *A Role for Sleep in Artificial Cognition through Deferred Restructuring of Experience in Autonomous Machines*. In: From Animals to Animats 13, Springer International Publishing, pp.1-10

[Lin2011] Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2014). Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, Mass.: The MIT Press.

[Wallach2008] Wallach, W., & Allen, C. (2008). Moral machines: Teaching robots right from wrong. Oxford University Press.

[Winfield2014] AFT Winfield, C Blum, W Liu (2014) *Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection*, Advances in Autonomous Robotics Systems, 85-96