Responsible Autonomy

Carles Sierra
IIIA-CSIC
Bellaterra, Barcelona, Catalonia
sierra@iiia.csic.es

ABSTRACT

The main challenge that artificial intelligence research is facing nowadays is how to guarantee the development of responsible technology. And, in particular, how to guarantee that autonomy is responsible. The social fears on the actions taken by AI can only be appeased by providing ethical certification and transparency of systems.[1] However, this is certainly not an easy task. As we very well know in the multiagent systems field, the prediction accuracy of system outcomes has limits as multiagent systems are actually examples of complex systems. And AI will be social, there will be thousands of AI systems interacting among themselves and with a multitude of humans; AI will necessarily be multiagent.

Although we cannot provide complete guarantees on outcomes, we must be able to define with accuracy what autonomous behaviour is acceptable (ethical), to provide repair methods for anomalous behaviour and to explain the rationale of AI decisions. Ideally, we should be able to guarantee responsible behaviour of individual AI systems *by construction*.

I understand by an ethical AI system one that is capable of deciding what are the most convenient norms, abide by them and make them evolve and adapt. The area of multiagent systems has developed a number of theoretical and practical tools that properly combined can provide a path to develop such systems, that is, provide means to build ethical-by-construction systems: agreement technologies to decide on acceptable ethical behaviour, normative frameworks to represent and reason on ethics, and electronic institutions to operationalise ethical interactions. Along my career I have contributed with tools on these three areas [1, 2, 5]. In this keynote I will describe a methodology to support their combination that incorporates some new ideas from law [3], and organisational theory [4].

KEYWORDS

Responsible AI; Agreement Technologies

Short Bio

Carles Sierra is a Research Professor of the Artificial Intelligence Research Institute (IIIA-CSIC) in the area of Barcelona. He is currently the Vice-Director of the Institute. He received his PhD in Computer Science from the Technical University of Barcelona (UPC) in 1989. He has been doing research on Artificial Intelligence topics since then. He has been visiting researcher at Queen Mary and Westfield College in London (1996-1997) and at the University of Technology in Sydney for extended periods between 2004 and 2012. He is also an Adjunct Professor of the Western Sydney University. He has taught postgraduate courses on different Ai topics at several Universities: Université Paris Descartes, University of Technology, Sydney, Universitat Politècnica de València, and Universitat Autònoma de Barcelona among others.

He has contributed to agent research in the areas of negotiation, argumentation-based negotiation, computational trust and reputation, team formation, and electronic institutions. These contributions have materialised in more than 300 scientific publications. His current focus of work gravitates around the use of AI techniques for Education and on social applications of AI.

Also, he has served the research community of MAS as General Chair of the AAMAS conference in 2009, Program Chair in 2004, and as Editor in Chief of the Journal of Autonomous Agents and Multiagent Systems (2014-2019). Also, he served the broader AI community as local chair of IJCAI 2011 in Barcelona and as Program Chair of IJCAI 2017 in Melbourne. He has been in the editorial board of nine journals. He has served as evaluator of numerous calls and reviewer of many projects of the EU research programs. He is an EurAI Fellow and was the President of the Catalan Association of AI between 1998-2002.

REFERENCES

[1] Adrian Perreau de Pinninck, Carles Sierra, and W. Marco Schorlemmer. 2010. A multiagent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems* 21, 3 (2010), 397–424. https://doi.org/10.1007/s10458-009-9107-8
[2] Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodríguez-Aguilar, and Carles Sierra. 2012. Communicating open systems. *Artif. Intell.* 186 (2012), 38–94. https://doi.org/10.1016/j.artint.2012.03.004
[3] Wesley Hohfeld. 1913. Some Fundamental Legal Conceptions as Applied in Legal Reasoning. *Yale Law Journal* 23 (1913), 16–59.
[4] Elinor Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press.
[5] Carles Sierra, Vicente J. Botti, and Sascha Ossowski. 2011. Agreement Computing. *KI* 25, 1 (2011), 57–61. https://doi.org/10.1007/s13218-010-0070-y

[1] See for instance the Barcelona declaration https://www.iiia.csic.es/barcelonadeclaration/

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 2019, Montreal, Canada