Autonomous Systems and Society

An ***Autonomous System*** (AS) is a piece of software that makes decisions and then acts upon them, in a way that has real world consequences. While autonomous systems often use Artificial Intelligence (AI) techniques, the two are distinct. For instance, a thermostat is autonomous, but unintelligent. On the other hand, an AI system acting solely as an advisor to a human does not act on its recommendations, and so is not autonomous.

The past few years have witnessed the rapid emergence of autonomous systems in our lives. Examples of autonomous systems include robots that support elderly people with dementia to live independent lives, smart homes, software assistants, Unmanned Aerial Vehicles, self-driving cars, and certain forms of manufacturing.

However, these systems raise many questions. Answering these question requires a range of disciplines (highlighted below). It also requires technological knowledge and credibility, in order to avoid falling into the trap of positing implausible technologies (e.g. "superintelligence").

The "agents" research group in the Information Science department has both substantial expertise (e.g. Padgham & Winikoff (2004)) and an international reputation in this area (e.g. Professor Winikoff has been both general and programme chair of the top international conference in the area), as well as local reputation (e.g. http://www.otago.ac.nz/hekitenga/2016/otago629405.html and http://www.otago.ac.nz/business/research/otago597062.html)

We group the many questions raised under two broad areas: the **acceptability** of the technology to society, and **how society should respond**.

Let us first consider **acceptability**. There are two key questions. The first is: *which domains of application are acceptable?* There are some applications where human involvement in decision making is essential. For example, an autonomous system that handed down prison sentences instead of a human judge may not be socially acceptable, even though human judges are known to be biased and to be affected by factors such as hunger (Danziger *et al.*, 2011). There is also a strong case for banning the development of autonomous weapons. The question of acceptable application domains is about ethics, and requires public discussion and engagement.

The second key question relating to acceptability concerns **trust**: *in what situations will humans come to (appropriately) trust autonomous systems?* Note the use of "appropriately" - we do not seek blind trust, but appropriate levels of trust. Note also the use of "come to" - building trust is a process.

We posit (Winikoff, 2017) that there are a number of prerequisite elements that must be present in order for appropriate trust to be formed. However, an overarching research question is to empirically explore the prerequisites to trust. Which of the prerequisites discussed below are actually required? How is this affected by other factors such as the application domain, and the consequences of the system making poor decisions? And are there additional prerequisites to trust?

This overarching question requires expertise from the disciplines (and departments) of Marketing and Management, and it follows similar methodologies to some of their existing related work on trust of self-driving cars (Wooliscroft, unpublished report for Ministry of Transport - a related presentation can be found at http://bit.ly/2pj3ZvQ) and on trust factors in smart homes and the Internet of Things (Harwood & Garry, 2017). However, in addition to studying trust as a *static* concept, we also need to consider the *process* of gaining (and reacquiring) trust over time (Grover *et al.*, 2014 & 2017).

We also posit that is it desirable for autonomous systems to have a representation of human values, and be able to use these in their codified reasoning processes (Cranefield *et al.*, 2017). Such values may include cultural expectations and norms (e.g. Tikanga Māori), and ethics. For example, consider a system that takes care of an aged person, perhaps with dementia or Alzheimer's disease. There are situations where competing options may be resolved by considering human values, such as autonomy vs. safety, or privacy vs. health. Perhaps the elderly person wants to go for a walk (which is both healthy, and is aligned with their desire for autonomy), but for safety reasons they should not be permitted to leave the house alone. The system needs to decide whether to allow the person it is caring for to leave the house, and, if so, what other actions may need to be taken. Developing representations for values and computational reasoning mechanisms requires expertise from Information Science, as well as from Marketing and Management to clarify the sorts of values and reasoning that are required to support trust.

We also posit that it is essential for autonomous systems to be able to *explain why* they made certain decisions. The explanation needs to be in a form that is comprehensible and accessible, and may be interactive (i.e. take the form of a dialogue, rather than a single query followed by a complex answer). Developing such mechanisms requires Information Science expertise to construct computational mechanisms, and Marketing and Management expertise to assess them, to provide feedback to refine iterated development.

Finally, we posit that a necessary prerequisite for trust is the existence of a societal mechanism that provides for *recourse* for someone who has been adversely affected by an autonomous system's poor decision. This could be legal, or a form of insurance (*a la* ACC www.acc.co.nz).

We now turn to the second broad area: how does society need to change in response to the development and deployment of autonomous systems?

This question can be considered from various perspectives. One important viewpoint is that of individual businesses (and other organisations). How does the deployment of autonomous systems affect such organisations? How does it affect the nature and future of work? The ongoing research project in the Management department on *Work Futures Otago: Trends, Disruptions and Transitions* by O'Kane, Ruwhiu, and Walton is relevant in that questions asked around Autonomous Systems are interlinked with the way that we work, and the broader society in which we operate. One key question this existing research focuses upon is *plausibility*, asking how we can think plausibly into the future. A second key question is which skills will be needed in the workforce in the future, and how can society plan now to ensure these skills are present? (O'Kane *et al.*, 2017). This line of enquiry also raises questions around the impact of autonomous systems on well-being in the workplace, which involves expertise from the Tourism department.

A second viewpoint is that of a society as a whole. What are the likely impacts of autonomous systems on the overall workforce in a country or region? This question is still not yet settled: it is clear that some occupations may not exist in decades to come (Autor, 2015), but the extent is unclear. It is also unclear to what extent jobs that are rendered redundant will be replaced with other jobs, as has been the case with previous technological disruptions (Brynjolfsson & McAfee, 2014), or whether this time is truly different. Answering this question involves expertise from the department of Economics , specifically labour economics.

In addition, there is also a need to examine the role of work in society, and the idea of a Universal Basic Income (UBI) (Van Parijs, 2004). Is the combination of autonomous systems and a UBI perhaps finally leading to a society with less work and more leisure time? These questions, which would benefit from considering an intergenerational collectivist perspective, are broad, and inherently multidisciplinary, requiring expertise from all of the departments and people mentioned above, and others.

Bibliography

D.H. Autor. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. Journal of Economic Perspectives 29(3), 3-30 (2015)

E. Brynjolfsson and A. McAfee. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W.W. Norton & Company (2014)

Stephen Cranefield, Michael Winikoff, Virginia Dignum, and Frank Dignum. No Pizza for you: Value-based plan selection in BDI agents. International Joint Conference on Artificial Intelligence, pages 178-184, doi:10.24963/ijcai.2017/26 (2017)

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. PNAS 2011 108 (17) 6889-6892 (2011), doi:10.1073/pnas.1018033108

Steven L. Grover, Marie-Aude Abid-Dupont, Caroline Manville, and Markus C. Hasel. Repairing Broken Trust Between Leaders and Followers: How Violation Characteristics Temper Apologies. Journal of Business Ethics, doi:10.1007/s10551-017-3509-3 (2017)

Steven L. Grover, Markus C. Hasel, Caroline Manville, and Carolina Serrano-Archimi. Follower reactions to leader trust violations: A grounded theory of violation types, likelihood of recovery, and recovery process, European Management Journal, 32(5):689-702, ISSN 0263-2373, doi:10.1016/j.emj.2014.01.002 (2014)

Tracy Harwood and Tony Garry. Internet of Things: understanding trust in techno-service systems. Journal of Service Management, 28(3):442-475, doi:10.1108/JOSM-11-2016-0299 (2017)

Wesley Kukard and Lincoln Wood. Consumers' perceptions of item-level RFID use in FMCG: A balanced perspective of benefits and risks. Journal of Global Information Management, 27(1), 21-42, doi:10.4018/JGIM.2017010102 (2017)

P. O'Kane, D. Ruwhiu, S. Walton, and F. Edgar. How Should we Respond? Worker Skill Development in 2040. 77th Annual Meeting of the Academy of Management, Atlanta, August (2017)

Lin Padgham and Michael Winikoff. Developing Intelligent Agent Systems: A Practical Guide. ISBN 0-470-86120-7, John Wiley and Sons (2004)

P. Van Parijs. Basic Income: A Simple and Powerful Idea for the Twenty-First Century. Politics & Society, 32(1), pp.7-39 (2004)

Michael Winikoff. Towards Trusting Autonomous Systems. Workshop on Engineering Multi-Agent Systems (EMAS) (2017).