Decision Making in Robots and Autonomous Agents: Suggested Topics for Term Paper (Semester 2 - 2017/18)

Subramanian Ramamoorthy

6 February 2018

1 Instructions

- The term paper is to be written and submitted before the end of the term. It will be the equivalent of a homework assignment 3, to be done either individually, or in groups of two (in which case, the expectation will be of a correspondingly more polished level of work). Without prejudicing your preference, we recommend team work.
- The primary aim is to give you the chance to go further in depth into one chosen topic of your choice, by summarising and critically evaluating one (or a few) key paper(s) in that sub-field.
- The outcomes will be twofold. Firstly, we will expect a written report, not longer than 4 pages in length, written in the IEEE conference paper format https://www.ieee.org/conferences_events/conferences/publishing/templates.html. Secondly, we will expect a presentation in class, for which the last two lecture hours of this term have been reserved.
- This assignment will count for 10% of your final course mark. The written report is due at 4 pm on 27 March 2018.

Good Scholarly Practice: Please remember the University requirement as regards all assessed work for credit. Details about this can be found at:

http://web.inf.ed.ac.uk/infweb/admin/policies/academic-misconduct

2 Topics

The list below is a suggested list of topics. We would expect that most students will (individually or in groups of two) select one of these topics, which roughly follows the structure of the course.

Students are welcome to propose other topics, by emailing the instructor and discussing it with him, in which case that can be added to this list.

As an initial guideline (to be elaborated in class), students should consider the following when researching the topic and writing the report.

A good report recapitulates the content of the paper without merely copying it. So, critical evaluation is key, as is explaining background that you or your peers may need to fully grasp the concept. If you were able to go further into the literature and give a broader perspective on the topic, that is a bonus, but the initial expectation is merely that you are reporting on the content at the level of a single sound paper or two.

A couple of the topics may build directly on material covered in class. In that case, the expectation is that your report and presentation takes the next step forward from where we left off.

1. *Encoding robot control behaviours in terms of hybrid systems.*
    - J. Pratt, C.-M. Chew, A. Torres, P. Dilworth, G. Pratt, Virtual Model Control: An intuitive approach for bipedal locomotion, Int. J. Robotics Research, Vol 20, Issue 2, pp. 129 - 143, 2001.
    - C. Belta, A Bicchi, M. Egerstedt, E. Frazzoli, E. Klavins, G. Pappas, Symbolic planning and control of robot motion [grand challenges of robotics], IEEE Robotics & Automation Magazine 14(1):61-70, 2007.
2. *Robot motion planning in the presence of humans.*
    - A.D. Dragan, Robot planning with mathematical models of human state and action. https://people.eecs.berkeley.edu/~anca/papers/summary.pdf.
3. *Decision theoretic methods for diagnosis.*
    - (classic - start here and find more modern links) E.J. Horvitz, J.S. Breese, M. Henrion, Decision theory in expert systems and artificial intelligence, International Journal of Approximate Reasoning, Volume 2, Issue 3, pp. 247-302, 1988.
4. *Decision theoretic methods for accident analysis.* (this is a diverse literature - view starting points below critically and explore)

- R. Barkan, D. Zohar, I. Erev, Accidents and decision making under uncertainty: A comparison of four models. Organizational behavior and human decision processes, 74(2), 118-144, 1998.
- S. Oppe, The concept of risk: A decision theoretic approach. Ergonomics, 31(4), 435-440, 1988.
- C. Perrow, Normal accidents: Living with high risk technologies. Princeton university press, 2011.

5. *Eliciting preferences regarding choice behaviour.*

   - D. Braziunas,C. Boutilier. Preference elicitation and generalized additive utility. In Proceedings of the National Conference on Artificial Intelligence, vol. 21, no. 2, p. 1573. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

6. *Decision theoretic methods for Human Computer Interactions.*

   - JR Hauser, GL Urban, G Liberali, M Braun, Website morphing. Marketing Science, 28(2), 202-223, 2009.

7. *Modelling location privacy.*

   - R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, J.-Y. Le Boudec. Protecting location privacy: optimal strategy against localization attacks. In Proc. ACM conference on Computer and communications security (CCS '12), 2012. http://dx.doi.org/10.1145/2382196.2382261.

8. *Responsibility and Blame through the notions of Actual Causality.*

   - JY Halpern, C Hitchcock, Actual causation and the art of modeling, . arXiv preprint https://arxiv.org/abs/1106.2652, 2011.
   - JY Halpern, Actual Causality, MIT Press, 2016.

9. *Robot safety 1 (Mobileye).*

   - S Shalev-Shwartz, S Shammah, A Shashua, On a formal model of safe and scalable self-driving cars, arXiv preprint, https://arxiv.org/pdf/1708.06374.pdf, 2017.

10. *Robot safety 2 (Berkeley and Google).*

- D. Hadfield-Menell, A.D. Dragan, P. Abbeel, S. J. Russell, The off-switch game, arXiv preprint, https://arxiv.org/abs/1611.08219, 2017.
- D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Man, Concrete Problems in AI Safety, arXiv preprint, https://arxiv.org/abs/1606.06565, 2017.

11. *Game theory based models of multi-agent interaction.*

    - JS Rosenschein, G Zlotkin, Designing conventions for automated negotiation, https://doi.org/10.1609/aimag.v15i3.1098, AAAI Magazine, 1994.
    - JS Rosenschein, G Zlotkin, Rules of Encounter: Designing Conventions for Automated Negotiation Among Computers, MIT Press, 1994.

12. *Nudging and Choice Architecture.*

    - R.H. Thaler, C.R. Sunstein, J.P. Balz, Choice architecture. The Behavioral Foundations of Public Policy, Ch. 25, Eldar Shafir, ed. (2012). http://dx.doi.org/10.2139/ssrn.2536504.
    - Johnson, E. J., Shu, S. B., Dellaert, B. G., Fox, C., Goldstein, D. G., Hubl, G., Wansink, B. Beyond nudges: Tools of a choice architecture. Marketing Letters, 23(2), 487-504, 2012.

13. Further investigations into explainable and interpretable Artificial Intelligence (contact instructor, if interested).