Agent Coordination and Potential Risks: Meaningful Environments for Evaluating Multiagent Systems

Nader Chmait and **David L. Dowe** and **David G. Green** and **Yuan-Fang Li**
Faculty of Information Technology, Clayton, Monash University, Melbourne, Vic. 3800, Australia
{nader.chmait, david.dowe, david.green, yuanfang.li}@monash.edu

Abstract

There are two major ways in which intelligence can emerge in distributed artificial systems: by design or, alternatively, due to simple ad-hoc interactions between the agents without prior coordination. At present, there is no clear measure of intelligence for state-of-the-art artificial agent behaviours operating in realistic multiagent scenarios. Therefore, possible emergent intelligence in such systems (and its potential benefits and risks) cannot be anticipated. This paper gives insights into some general properties that make a testing environment feasible for measuring intelligence in multiagent systems and monitoring their (collective) behaviour. We discuss some evaluation guidelines and present a preliminary methodology for evaluating coordination between interactive agents, and assessing risks that can potentially arise from multiagent interactions.

1 Artificial Intelligence and Potential Risks: A Brief Overview

It is known that interaction between agents (whether human, animal or artificial) can improve their overall performance and create intelligent collectives. However, phenomena like the tragedy of the commons [Hardin, 1968] and the prisoners' dilemma [Rapoport and Chammah, 1965] can arise in systems or societies of rational agents, where each individual behaving by doing its best can have undesirable - and sometimes seriously adverse - implications on safety and resource management. These problematic scenarios occur between different types of agents, whether human or artificial. Examples include the 2008 US housing crisis [Kotz, 2009] and climate change, which are cases of the tragedy of the commons.

Distributed artificial agents, like human agents, also (usually) hold a common pool of resources and have their own policies that determine how to use these resources, which can subsequently lead to similar adverse scenarios such as the ones described earlier [Turner, 1993]. Other important multi-agent environment settings that are ubiquitous in many online and web applications must be rapidly investigated, especially those settings in which artificial agents share and delegate rewards which are basically two fundamental models of how agents are motivated to collaborate. For instance, machine learning algorithms are extensively used nowadays in many business and management operations. Given that such algorithms can share or delegate rewards, it is becoming increasingly complex for human agents to directly control, or even predict, what is going to happen in an environment in which nearly all productive work, including management, investment, and the design of new machines, is being handled by artificial agents. In fact, nowadays coordination is inherent in an enormous range of research disciplines and applications from robotics to complex adaptive systems to human-computer interaction and every-day human interactions. The Internet of Things (IoT) is a good example of the rapid spread of coordination between agents (via machine-to-machine and machine-to-human communication) in the aim of building smart grids, homes, cities, etc.

Bostrom outlined in his book on Superintelligence [Bostrom, 2014] many future risks that can arise from intelligent AI agents that can potentially present dire challenges to humans. Examples of such include artificial agents misidentifying their objectives, resistance to change goals, instrumentality (e.g., humans as resources and the paperclip AI problem) and, more importantly, *unpredictability* where intelligent AI agents cannot be certified for unseen situations. Solomonoff [Solomonoff, 1967; Dowe, 2013] also warned of the dangers of a very intelligent machine and discussed in [Solomonoff, 1985] six future milestones in AI, ranging from "the development of a very general theory of problem solving to the creation of machines with capacities well beyond those of a single human".

AI risks are not confined to the future. Financial disasters have already occurred due to disruptive and deceitful artificial agent behaviours (purposely designed by humans) such as, for example, "the 2010 Crash of 2:45" [Kirilenko *et al.*, 2017] which was described as one of the most turbulent periods in the history of financial markets. Moreover, prominent intellectuals such as Stephen Hawking and those affiliated with the likes of the Future of Humanity and the Future of Life institutes have repeatedly warned about concrete and prominent problems (e.g., see [Dowd, 2017]) of AI such as machine learning-based systems controlling industrial processes, health-related systems, and other mission-critical technology [Amodei *et al.*, 2016]. At the moment, it is still unclear whether many of these issues originate from too little or too much machine/artificial intelligence, or alternatively as a result of some emergent phenomena from the interaction of more than one agent.

In this paper, we give insight into some general features that we believe to be important to make an *environment* meaningful or feasible for measuring *intelligence* in a coordinated multiagent systems. The next section gives a brief background on some of the intelligence test platforms that have been recently used to evaluate AI agents. We then present a few guidelines for evaluating coordination between interactive agents and disclosing some of the potential risks arising from such interactions. We touch upon some of the argued ideas relating to AI and safety which might become more and more relevant along the path to the technological singularity. Finally, we conclude with a short summary and give directions for future work.

2 Assessment Tasks and Environments

There are two major ways in which intelligence can emerge in multiagent systems. The first one is by design, where agents are put together in predefined settings in such a way to increase their performance or world utility function. The second way is when intelligence emerges in the system due to simple ad-hoc interactions between the distributed agents, and thus without prior coordination. Despite the latest research milestones in artificial intelligence, there is no clear measure of intelligence for state-of-the-art AI agent behaviours operating in realistic multiagent scenarios, especially with respect to problems involving hierarchical tasks requiring some sort of coordination between the agents. Therefore, possible emergent intelligence in such systems, and its potential benefits and risks, cannot be anticipated. The focus on multiagent systems and coordinated agents in particular is mainly due to their unpredictability. For instance, it is usually much harder to understand what is going on in systems of agents (e.g., distributed artificial agents, systems of neurons making up a human brain, swarms, etc.) than with one agent. Moreover, it is natural that, under certain circumstances, systems with more and more parts to them might become increasingly unstable. Given the very sophisticated and unpredictable behaviour that can emerge from multiagent coordination, we argue that a methodology for assessing coordination abilities in these systems is inevitable if one wants to anticipate or gain insight into their potential impact - whether it is positive or negative.

Many intelligence tests are designed to evaluate individual agents, and thus (unless they can be extended to multiagent scenarios, they) can be excluded as appropriate for achieving the objectives outlined in this paper. Those tests that do allow for the evaluation of multiagent systems have limitations. For instance, many of these tests usually do not take into consideration agent coordination as part of the assessment but rather simply return an average performance of a group of evaluated agents over a set of general evaluation tasks (e.g., regarding compression [Dowe and Hajek, 1997a; Dowe and Hajek, 1997b; Dowe and Hajek, 1998; Hernández-Orallo and Minaya-Collado, 1998; Mahoney, 1999; Hernández-Orallo, 2000], and others [Sanghi and Dowe, 2003]). Recently, [Hernández-Orallo *et al.*, 2017] reported on a series of new platforms and events dealing with AI evaluation, in particular those that may change the way in which AI systems are compared and how their progress is measured. One such platform is Microsoft's Project Malmo [Johnson *et al.*, 2016], which has been looking into the evaluation of collaborative tasks for AI agents. The project presents a wide range of experimentation scenarios for evaluating reinforcement learning agents and general AI research over a tasks ranging from navigation and survival to collaboration and problem solving. Other interesting evaluation platforms like OpenAI Gym [Brockman *et al.*, 2016] and Facebook's TorchCraft [Synnaeve *et al.*, 2016] provide environments in which once can evaluate machine learning learning agents over diverse collections of (reinforcement learning) tasks including some that require coordination. We have also discussed in our earlier work on collective (artificial) intelligence [Chmait *et al.*, 2016a; Chmait *et al.*, 2016b; Chmait *et al.*, 2015], a simple dynamic interactive setting for measuring the performance of cooperative artificial agents interacting under various cooperation strategies and group organisational (or network) structures, and touched upon how studies of intelligence might connect to different research areas such as business decision-making [Chmait, 2017]. For a thorough historical background on the evaluation of intelligence of various cognitive systems (including machines) refer to [Hernández-Orallo, 2017], which provides an integrated view of the evaluation of natural and artificial intelligence.

The above-described environments are feasible for assessing state-of-the-art (machine learning) agent performances and perhaps evaluating their social abilities, but little do they tell us about how well these agents can (specifically) coordinate and the risks that might (or might not) occur as a result of their (collective) behaviour/operation. Thus, one key objective is to try to identify multiagent coordination scenarios that might result in (realistic) undesirable consequences, and understand how to measure their impact (among other things) on real world applications.

3 Desired Features For Evaluation of Distributed Agents and Risks

In order to address the objective just mentioned in the previous paragraph, the first step is to create a set of testing environments with some underlying desired features that make it possible to:

1. Evaluate *when* and *how* coordination can arise from (ad-hoc) agent interactions.

2. Monitor the performance of groups of interactive agents over problems specifically requiring coordination. These problems could range from very simplistic tasks, such as (two or more) robots lifting and moving a table, to more sophisticated multiagent settings in which the agents are engaged in $N - person$ prisoners' dilemma [Colman, 2014, Sec. 8].

3. Identify the factors that influence coordination in agent collectives.

4. Identify potential risks arising from (rational) agent coordination and the situations that might have led to such risks (e.g., multiagent configuration settings or unexpected changes in the environment). This includes risks arising from human-machine interaction or any type of undesirable circumstances from resource depletion, to safety and denial of service, etc. Risks can further be classified into different categories corresponding to their emergency, and urgency of their consequences (e.g., their timeliness, urgency of being addressed in the near future, short term vs. long term consequences).

We define a *risk* to be any circumstance of non-zero probability which can adversely affect utility. This includes (e.g.) a decrease in a quantified utility function or even simply an inferior location in a (possibly unquantified) *partial order* of utilities. As a comparatively simple case in point, we consider the prisoners' dilemma [Rapoport and Chammah, 1965] where (e.g.) the prisoners get joint utility $(-9, -9)$ by taking actions leading to a Nash equilibrium [Maskin, 1999] whereas the curiously *individually sub-optimal* (or less secure) strategy of cooperation leads to the outcome $(-1, -1)$ with all parties better off.

Thus, by developing a proper intelligence test framework over which interactive agents can be assessed collectively, on well-defined tasks that *specifically require coordination* to be solved, not only can we benefit from measuring the coordination skills of the agents, but also we can detect some of the adverse risks that might arise from such types of coordination. In other words, a measurement technique for the quantitative assessment of coordination between intelligent distributed agents can be employed to get insights into the risks that can arise from the interactions of such agents. We develop these ideas in the next section.

4 Evaluating Coordination and Disclosing Risks: A Preliminary Methodology

We present a few ideas on how to evaluate the difficulty of problems requiring multiagent coordination over a collection of subtasks. Based on (algorithmic) information theory (AIT) and (Solomonoff-)Kolmogorov complexity [Solomonoff, 1964a; Solomonoff, 1964b; Li and Vitányi, 2008], LNPPP [Dowe, 2008, Sec. 0.2.7, 1st bullet point][Dowe, 2011, Sec. 5.3, p. 936][Dowe, 2013, Sec. 4.7, p. 24] is a method proposed to give universal distributions over (environments of) statistical and machine learning problems to compare the efficacy of rival estimators (e.g., AIC vs BIC). In essence, LNPPP compares - and ranks - a weighted sum of the respective penalties (e.g., squared error). In similar vein, AIT and Kolmogorov complexity have been used to measure environment complexity [Legg and Hutter, 2007; Hernández-Orallo and Dowe, 2010] and task difficulty [Hernández-Orallo, 2017]. For more realistic evaluation over real-world agents in real-world environments and the relevant spatio-temporal considerations, it appears appropriate to consider issues of (redundant TMs and) resolution [Dowe, 2008, Sec. 0.2.7, p. 544, col. 2][Hernández-Orallo and Dowe, 2010, p. 1514, footnote 6][Dowe, 2013, sec. 4.4][Dowe and Hernández-Orallo, 2014, Sec. 5 and elsewhere].

Earlier work on the performance of multiagent system [Chmait *et al.*, 2016a; Chmait *et al.*, 2015] showed that groups do not always outperform the same selection of agents working in isolation as commonly presumed. Coordination was found to be a major factor among others controlling the performance of these groups. Many tasks and problems that require coordination can be used to evaluate a group of subjects. Nevertheless, quantitatively measuring coordination between interactive agents can be a very difficult task for several reasons. For instance, formalising the assessment tasks or problems that require coordination is not trivial. While humans and some non-human animals (e.g., apes and elephants) possess mirror self-recognition (MSR) abilities [Plotnik *et al.*, 2006] which are partially responsible for their complex sociality and cooperation skills, in many cases, artificial agents need to be fed information about their environment in advance using advanced knowledge representation techniques.

In order to design and implement a practical new intelligence test framework that will particularly allow for the measurement of coordination between various types of agents, important problems in AI that require coordination must be identified and presented to interactive artificial agents in the form of intelligence tests, in such a way that payoff only occurs if two or more agents perform a particular sequence of actions adhering to some coordination scheme that is necessary to solve the task. To further assess the risks resulting from agents actions, the environment in which they operate should also be designed to respond to their behaviour and actions (e.g., resources might diminish, parameters of the environment could vary according to the sequence of actions of the agents). In addition, testing can be performed in an even more heterogeneous setting where the agents don't have a common reward function and/or actions and observations of their environments.

We propose a preliminary methodology to achieve the above objectives. The first step is to outline the set of important multiagent problems requiring coordination that can realistically or theoretically be performed by a group of interactive AI agents. We denote this set by P_{all}. Assuming that we have identified this set of meaningful problems, we randomly sample a subset $\mathcal{P} \subset P_{all}$, and proceed by analysing the hypothetical difficulties of the problems in $\mathcal{P}$, where a problem $p \in \mathcal{P}$ is a collections of $n > 1$ subtasks $\{t_1, t_2, \ldots, t_n\}$ to be solved via multiagent coordination. An example is given in Figure 1. The problem difficulty, which is categorised into four difficulty levels: very low, low, high and very high, is a function of the complexity of the tasks incorporated in that problem[1]. The number of such tasks is depicted on the y-axis

[1] Each data point appearing in Figure 1 (top) corresponds to a problem, which in turn is a collection of tasks that are to be solved via multiagent coordination. There are 10000 synthetic data points randomly generated using the Normal distribution with mean μ equal to 14 and a standard deviation σ of 3. For simplicity, the problem difficulty was calculated as the product of the total number of tasks in the problem and their maximum task complexity. However, for real data, the problem difficulty should rather be a function of its underlying tasks and their *individual* complexities. The difficulties were further categorised into four groups using different thresholds. For instance, very low, low, high and very high difficulty

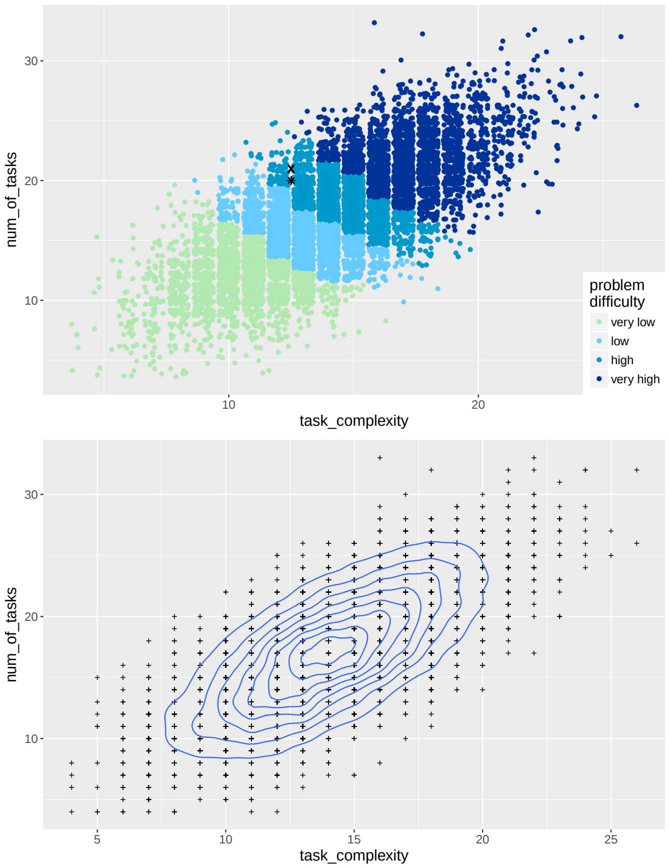

Figure 1: The top plot shows the hypothetical difficulties of (a sample drawn from the set of) synthetic important problems in AI measured as a function of the number of tasks incorporated in the problems (y-axis), and their theoretical complexities (x-axis) [Refer to footnote 1]. The Kernel Density Estimation (KDE) (or the probability density function estimate) of such problems is also illustrated in the bottom plot, showing higher concentration for problems consisting of 16 to 18 tasks with complexities within the range $[13, 15]$.

of Figure 1 (top), while the bound on the tasks' individual theoretical complexities is represented on the x-axis of the same plot. For example, the problem X that is depicted in Figure 1 (top) corresponds to a collection of $n = 20$ tasks (requiring multiagent coordination to be solved) of complexities less than or equal to (a hypothetical value of) 13, assuming that we have a flexible complexity scale with which we can formally assign complexity values to different types of tasks. It is reasonable to presume that the problem difficulty correlates with its underlying number of tasks and their complexity. One could further analyse the densities (or the probability density function estimate) of such problems in order to identify the most frequent ones, and perhaps categorise them according to their timeliness and the type of tasks they require to be solved, etc. Subsequently, intelligence tests would be developed to assess different types of interactive artificial (and hybrid human-machine) agent groups over a similar set and number of tasks requiring coordination. The behaviour and performance of the evaluated agents is monitored along with other relevant factors such as: the consumed resources within a period of time, the disclosure of conflicts in decision-making (especially between human and AI agents) and coalition formation based on different opinions, along with many other safety-related factors (e.g., blocking rewards from, and taking adverse actions against, other agents sharing the environment).

Moreover, considering the two different ways that can lead to multiagent coordination, the same groups of agents are tested in two different settings where, in the first one, they are designed to coordinate in order to solve the problem tasks whereas, in the second one, they are monitored across various group organisational structures to detect the presence of ad-hoc coordination and unexpected collaborative interactions. Finally, all undesirable scenarios recorded from the previous simulations are analysed and linked to real-world applications/scenarios in which they might occur. Further efforts are required to devise new methodologies to overcome and avert such scenarios.

In fact, some multiagent approaches to avoid undesirable effects (e.g., make sure a human is not blocked by an agent from shutting the agent down) were discussed in [Amodei *et al.*, 2016, Secs. 3 and 4] such as the Cooperative Inverse Reinforcement Learning [Hadfield-Menell *et al.*, 2016], and avoiding reward hacking. These can be thought of as some of the example problems illustrated in Figure 1 (top), in which case the tasks presented to the agents are well-known and properly defined in the purpose of monitoring a particular aspect of their behaviour (e.g., hacking or blocking rewards). Identifying and averting the underlying risks in these scenarios is relatively easier than in the case where agents might have developed ad-hoc coordination strategies, which can lead to unexpected outcomes.

5 Along the Path to the Technological Singularity

Not all speculations concerning superintelligent machines are pessimistic [Good, 1965]. In the last decade, research on AI safety has become popular among academic researchers and professionals and within various industries investing in state-of-the-art technologies and computer applications such as Google, Facebook and the Future of Life Institute. Consequently, debates on the ethical and societal implications of artificial intelligence have taken place inside these communities, and have also been of interest to the media [Dowd, 2017; Dowe, 2014].

However, these ideas are not new. They have been around for a long time from computer science research [Solomonoff, 1967; Solomonoff, 1985] to fiction novels [Asimov, 1950; Lem, 1964]. The rapid advance in technological innovations, and the increase in responsibility delegation to more adept machines, will likely lead to mainstream societal implica-

problems are respectively those with difficulties that are (i) within the 1st quantile of the (difficulty) distribution, (ii) between the 1st quantile and the median, (iii) between the median and the 3rd quantile and (iv) beyond the 3rd quantile.

tions in the near future and hence must be seriously investigated a priori. In fact, even if one is not particularly a fan of the idea of singularity, and the existential threat that AI might pose one day to humanity[2], it is irrational to deny risks that might result from the delegation of responsibility. While some of these responsibilities are routine mathematical or automated operations (e.g., calculators, small-scale automated manufacturing), others like driver-less cars, air traffic control systems and guided missiles [Dowe, 2014] are a lot more serious and there is much more at stake if such systems fail or malfunction. Moreover, if they do fail leading to (human) casualties, financial disasters and so on, an ensuing challenge is how to determine who takes the responsibility. Likewise, other problems resulting from the delegation of responsibility include unemployment, which directly relates to human social status [Graetz, 2015], and raises many questions regarding the future of employment [Frey and Osborne, 2017; Levy and Murnane, 2012]. For instance, AI might well push forward the idea of implementing a Universal Basic Income (UBI) [De Wispelaere and Stirton, 2004]. This has already triggered some science-fiction-like questions such as "should robots pay tax?" [Abbott and Bogenschneider, 2017], or have rights [McNally and Inayatullah, 1988]. Finally, even faultless AI systems can result in undesirable outcomes. For example, in the medical field, self-taught AI and machine learning techniques are being used for diagnostics and help in identifying adequate types of treatments for different patients, in many cases outperforming doctors [Weng *et al.*, 2017]. However, if physicians increasingly adopt and delegate responsibility to such machine-learning methods, AI systems might make expert decisions and take actions on our behalf that are only locally beneficial but which have unpleasant consequences on human well-being.

6 Conclusion and Future Work

We have discussed a preliminary methodology for the measurement of coordination in multiagent systems and its use for disclosing risks that might arise from their interactions. In brief terms, this can be done by first identifying a set of coordination problems that are particularly relevant to AI (and hybrid, human-artificial) agents, and then monitoring the agents' behaviours over these problems using an intelligence test framework. Two main scenarios were taken into consideration, where agents are designed to cooperate to solve problems and, alternatively, where coordination arises from the simple ad-hoc interactions between these agents. We gave insights into some general features that make testing environments feasible for evaluating multiagent systems and discussed some ideas that might be useful for assigning difficulty measures over coordination problems.

The risks described in this paper belong to a very broad scope. An important part of our future work is to provide a formal definition of what risk is, which could be in terms of the environment resources, the priority of decision-making and conflict resolution, partial ordering of utilities, etc. Moreover, a natural extension of this work is to devise a rigorous methodology for assessing the seriousness of these risks and provide (design and implementation) potential solutions in response.

References

[Abbott and Bogenschneider, 2017] Ryan Abbott and Bret N Bogenschneider. Should robots pay taxes? tax policy in the age of automation. *Harvard Law & Policy Review, Forthcoming*, 2017.

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. 2016. *arXiv preprint arXiv:1606.06565*.

[Asimov, 1950] Isaac Asimov. *I, Robot*. Gnome Press, New York, first edition, 1950.

[Bostrom, 2014] Nick Bostrom. *Superintelligence: Paths, dangers, strategies*. Oxford University Press, Oxford, UK, 2014.

[Brockman *et al.*, 2016] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. 2016. *arXiv preprint arXiv:1606.01540*.

[Chmait *et al.*, 2015] Nader Chmait, David L. Dowe, David G. Green, and Yuan-Fang Li. Observation, communication and intelligence in agent-based systems. In Jordi Bieger, Ben Goertzel, and Alexey Potapov, editors, *Proceedings 8th International Conference on Artificial General Intelligence*, volume 9205 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 50–59, Berlin, Germany, Jul 2015. Springer. `http://dx.doi.org/10.1007/978-3-319-21365-1_6`.

[Chmait *et al.*, 2016a] Nader Chmait, David L. Dowe, Yuan-Fang Li, David G. Green, and Javier Insa-Cabrera. Factors of collective intelligence: How smart are agent collectives? In Gal A. Kaminka, Maria Fox, Paolo Bouquet, Eyke Hüllermeier, Virginia Dignum, Frank Dignum, and Frank van Harmelen, editors, *Proceedings of 22nd European Conference on Artificial Intelligence ECAI*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 542–550, The Hague, The Netherlands, 2016. IOS Press. `http://ebooks.iospress.nl/volumearticle/44798`.

[Chmait *et al.*, 2016b] Nader Chmait, Yuan-Fang Li, David L. Dowe, and David G. Green. A dynamic intelligence test framework for evaluating AI agents. In *Proceedings of 1st International Workshop on Evaluating General-Purpose AI (EGPAI 2016), European Conference on Artificial Intelligence (ECAI 2016)*, pages 1–8, The Hague, The Netherlands, 2016. `http://www.ecai2016.org/content/uploads/2016/08/W14-EGPAI-2016.pdf`.

[2] One of the main attempts to provide an explanation to the *Fermi paradox* (the lack of evidence of other civilisations despite their high probability estimate of existence) is the theory that advanced civilisations are likely to have annihilated themselves as a result of wars, depletion of resources or malevolent artificial intelligence [Webb, 2002], such as killer robots.

[Chmait, 2017] Nader Chmait. Understanding and measuring collective intelligence across different cognitive systems (extended abstract). In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI 2017 Doctoral Consortium*, Melbourne, Australia, 2017. To appear.

[Colman, 2014] Andrew M Colman. *Game theory and experimental games: The study of strategic interaction*, volume 4 of *International Series in Experimental Social Psychology. Colman, Andrew M. (Eds)*. Pergamon, 2014.

[De Wispelaere and Stirton, 2004] Jurgen De Wispelaere and Lindsay Stirton. The many faces of universal basic income. *The Political Quarterly*, 75(3):266–274, 2004.

[Dowd, 2017] Maureen Dowd. Elon Musk's billion dollar crusade to stop the AI apocalypse. *Vanity Fair*, April 2017. `http://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x`.

[Dowe and Hajek, 1997a] David L. Dowe and Alan R. Hajek. A computational extension to the Turing Test. In *Proceedings of the 4th Conference of the Australasian Cognitive Science Society, University of Newcastle, NSW, Australia*, volume 1. Citeseer, 1997.

[Dowe and Hajek, 1997b] David L. Dowe and Alan R. Hajek. A computational extension to the Turing Test. Technical Report #97/322, Department of Computer Science, Monash University, Melbourne, Australia, 1997.

[Dowe and Hajek, 1998] D. L. Dowe and A. R. Hajek. A non-behavioural, computational extension to the Turing Test. In *International conference on computational intelligence & multimedia applications (ICCIMA'98), Gippsland, Australia*, pages 101–106, 1998.

[Dowe and Hernández-Orallo, 2014] David L Dowe and José Hernández-Orallo. How universal can an intelligence test be? *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, 22(1):51–69, February 2014. Sage Publications, Inc.

[Dowe, 2008] David L. Dowe. Foreword re C. S. Wallace. *The Computer Journal*, 51(5):523–560, 2008. Christopher Stewart WALLACE (1933-2004) memorial special issue.

[Dowe, 2011] David L. Dowe. MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness. In P. S. Bandyopadhyay and M. R. Forster, editor, *Handbook of the Philosophy of Science*, volume 7 of *Philosophy of Statistics*, pages 901–982. Elsevier, 2011.

[Dowe, 2013] David L. Dowe. Introduction to Ray Solomonoff 85th memorial conference. In David L. Dowe, editor, *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, volume 7070 of *Lecture Notes in Computer Science*, pages 1–36. Springer, Berlin, Heidelberg, 2013.

[Dowe, 2014] David L. Dowe. Is Stephen Hawking right? Could AI lead to the end of humankind? *The Conversation*, December 2014. `http://theconversation.com/is-stephen-hawking-right-could-ai-lead-to-the-end-of-humankind-34967`.

[Frey and Osborne, 2017] Carl Benedikt Frey and Michael A Osborne. The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, 114:254–280, 2017. Elsevier.

[Good, 1965] I. J. Good. Speculations concerning the first ultraintelligent machine. In F. Alt and M. Ruminoff, editors, *Advances in Computers, volume 6*. Academic Press, 1965.

[Graetz, 2015] Georg Graetz. Rise of the machines: The effects of labor-saving innovations on jobs and wages. *Centre for Economic Performance, London School of Economics and Political Science*, 2015.

[Hadfield-Menell *et al.*, 2016] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3909–3917, 2016.

[Hardin, 1968] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. American Association for the Advancement of Science.

[Hernández-Orallo and Dowe, 2010] José Hernández-Orallo and David L. Dowe. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539, December 2010. Elsevier Science Publishers Ltd.

[Hernández-Orallo and Minaya-Collado, 1998] José Hernández-Orallo and Neus Minaya-Collado. A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In *Proceedings of the Int. Symposium of EIS*, pages 146–163. ICSC Press, 1998.

[Hernández-Orallo *et al.*, 2017] Jose Hernández-Orallo, Marco Baroni, Jordi Bieger, Nader Chmait, David L. Dowe, Katja Hofmann, Fernando Martínez-Plumed, Claes Strannegård, and Kristinn R. Thórisson. A New AI Evaluation Cosmos: Ready to Play the Game? *Accepted, to appear in the AI Magazine, Association for the Advancement of Artificial Intelligence*, 2017.

[Hernández-Orallo, 2000] José Hernández-Orallo. Beyond the Turing Test. *Journal of Logic, Language and Information*, 9(4):447–466, October 2000. Springer.

[Hernández-Orallo, 2017] José Hernández-Orallo. *The measure of all minds: evaluating natural and artificial intelligence*. Cambridge University Press, 2017.

[Johnson *et al.*, 2016] Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The Malmo platform for artificial intelligence experimentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4246–4247, 2016.

[Kirilenko *et al.*, 2017] Andrei A Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High frequency trading in an electronic market. Journal of Finance (Forthcoming), 2017. `https://ssrn.com/abstract=1686004`.

[Kotz, 2009] David M Kotz. The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism. *Review of Radical Political Economics*, 41(3):305–317, 2009. SAGE Publications Sage CA: Los Angeles, CA.

[Legg and Hutter, 2007] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

[Lem, 1964] Stanislaw Lem. *Fables for Robots (short stories)*. Wydawnictwo Literackie, first edition, 1964.

[Levy and Murnane, 2012] Frank Levy and Richard J Murnane. *The new division of labor: How computers are creating the next job market*. Princeton University Press, 2012.

[Li and Vitányi, 2008] Ming Li and Paul Vitányi. *An introduction to Kolmogorov complexity and its applications (3rd ed.)*. Springer-Verlag New York, Inc., 2008.

[Mahoney, 1999] Matthew V Mahoney. Text compression as a test for artificial intelligence. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-99)*, page 970. John Wiley & Sons Ltd, 1999.

[Maskin, 1999] Eric Maskin. Nash equilibrium and welfare optimality. *The Review of Economic Studies*, 66(1):23–38, 1999. Oxford University Press.

[McNally and Inayatullah, 1988] Phil McNally and Sohail Inayatullah. The rights of robots: Technology, culture and law in the 21st century. *Futures*, 20(2):119–136, 1988. Elsevier.

[Plotnik *et al.*, 2006] Joshua M Plotnik, Frans BM De Waal, and Diana Reiss. Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences*, 103(45):17053–17057, 2006.

[Rapoport and Chammah, 1965] Anatol Rapoport and Albert M Chammah. *Prisoner's dilemma: A study in conflict and cooperation*, volume 165. University of Michigan press, 1965.

[Sanghi and Dowe, 2003] Pritika Sanghi and David L. Dowe. A computer program capable of passing I.Q. tests. In P. P. Slezak, editor, *Proceedings of the Joint International Conference on Cognitive Science, 4th ICCS International Conference on Cognitive Science & 7th ASCS Australasian Society for Cognitive Science (ICCS/ASCS-2003)*, pages 570–575, Sydney, NSW, Australia, 13-17 July 2003.

[Solomonoff, 1964a] Ray J Solomonoff. A formal theory of inductive inference. part I. *Information and Control*, 7(1):1–22, 1964. Elsevier.

[Solomonoff, 1964b] Ray J Solomonoff. A formal theory of inductive inference. part II. *Information and Control*, 7(2):224–254, 1964. Elsevier.

[Solomonoff, 1967] Ray J Solomonoff. Inductive inference research: status, Spring 1967, 1967. RTB 154, Rockford Research, Inc., 140 1/2 Mt. Auburn St., Cambridge, Mass. 0213.

[Solomonoff, 1985] Ray J Solomonoff. The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, 5(2):149–153, 1985. IOS Press.

[Synnaeve *et al.*, 2016] Gabriel Synnaeve, Nantas Nardelli, Alex Auvolat, Soumith Chintala, Timothée Lacroix, Zeming Lin, Florian Richoux, and Nicolas Usunier. Torchcraft: a library for machine learning research on real-time strategy games. 2016. *arXiv preprint arXiv:1611.00625*.

[Turner, 1993] Roy M Turner. *The tragedy of the commons and distributed AI systems*. Department of Computer Science, University of New Hampshire, 1993.

[Webb, 2002] Stephen Webb. *If the Universe Is Teeming with Aliens ... WHERE IS EVERYBODY?: Fifty Solutions to the Fermi Paradox and the Problem of Extraterrestrial Life*. Copernicus Series. Springer New York, 2002.

[Weng *et al.*, 2017] Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.