VICTOR GREEN & COMPANY
PATENT & TECHNICAL INFORMATION
Over 30 years of searching excellence

Asimov and the Patenting of Ethical Robots

January 2nd, 2020 marks the centenary of the birth of Isaac Asimov, the prolific science fiction writer known to millions as the author of the Foundation stories of Galactic Empire and for his statement of the Three Laws of Robotics, which in various forms inspired dozens of his robot stories, and influenced the work of many other writers and film-makers.

by Victor Green and Kaj Mahendiran
Victor Green & Company - Patent Information Analysts
www.victorgreen.co.uk

The Three Laws of Robotics are:

First Law

A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law

A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law

A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. [1]

It is not difficult to appreciate the simple power of this hierarchy of laws, nor the huge ambiguity they allow. For example, in the Second Law, which humans are to be obeyed? The fertility of this formulation and the variations it has engendered have proved inexhaustible.

As patent searchers, we were interested to see the extent to which the Three Laws and other ethical considerations appear in the patent literature on robotics and artificial intelligence (AI). We appreciated that a great deal of the intellectual property in the ethics of AI would reside in copyright rather than in patents, but thought the exploration on the patent side might provide some information not to be found elsewhere.

Method of search

The search aimed to find patents that fell under both of the search concepts AI (robots, artificial intelligence, machine learning, etc) and ETHICS. As the ETHICS concept would almost certainly have to be specified by keywords, we elected to combine it with patent classes for the AI concept rather than with keywords, so as to be bound within a high relevance context. The two Cooperation Patent Classification (CPC) classes defined below seemed a good place to start:

1. Artificial life, i.e., computers simulating life
2. Based on physical entities controlled by simulated intelligence to replicate intelligent life forms, e.g., robots replicating pets or humans in their appearance or behaviour

We used about 50 different expressions to capture the ETHICS concept. MORAL* and ETHIC*, of course, but ranging further from ASHAMED and ASIMOV via LETHAL* and OBLIGATION* to UNINTENDED CONSEQUENCES and WAR.

The retrieval was then examined to find candidates for discussion and analysis and to use them to identify further patent classification terms of interest. Five more classes were chosen for combination with the ETHICS concept that can be summarised as:

3. Programme-controlled manipulators, characterised by learning, adaptive, model-based, or rule-based expert control
4. Robotics programme-control systems which learn by operator observation, symbiosis, showing or watching
5. Control of position, course or altitude of land, water, air, or space vehicles, e.g., automatic pilot, characterised by the autonomous decision-making process, e.g., artificial intelligence, predefined behaviours
6. Computer systems based on simulated virtual individual or collective life forms, e.g., single avatar, social simulations, virtual worlds or particle swarm optimisation
7. Computer systems using neural network models

These do not, of course, exhaust the classes that might profitably be used in such a combination.

Some 550 patents were retrieved, of which about 50 were considered relevant for the subject. (This is a fairly typical relevance/retrieval ratio for a landscape using our approach to searching).

The Three Laws and AI

We reviewed these cases with the aim of identifying the main themes occurring in the patent literature, and in particular, the extent to which one or more of the Three Laws might be applicable to the inventions.

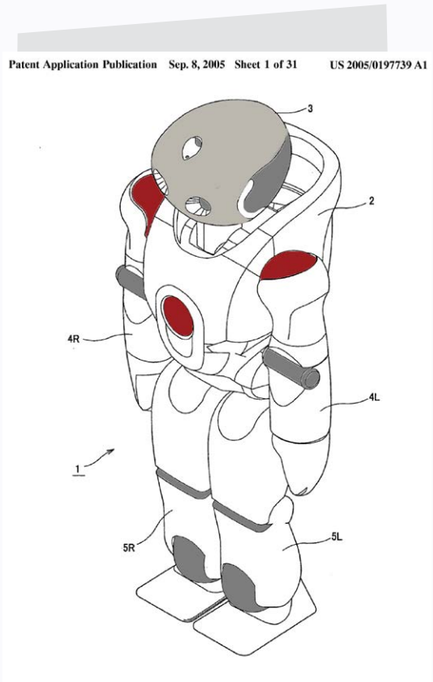

Fig. 1 Behaviour controlling system and behaviour controlling method for robot, US2005197739, Sony, September 8th, 2005

Patents are filed to protect rights that may be commercially or industrially valuable, so perhaps there should be no surprise to find so many concerned with the behaviour of autonomously driven road vehicles, a growing, and potentially transformative industry. These amounted to one-third of the selection. One recent application that took the eye, US 2019179306 assigned to Evexia Research, aims to prevent a driver using a vehicle as a weapon against pedestrians by simultaneously interpreting acoustic and accelerometer signals to detect actions such as high speed kerb-mounting, following which an unmodifiable algorithm brings the vehicle to the safest stop available, raises the windows, locks the doors and disables the accelerator.

Another five or so were also dedicated to specific applications such as prosthetic limbs preventing motions harmful to the patient's body, home robots acting in relatively uncontrolled spaces, a global telecommunication system, and a system for managing training programmes that simulate human behaviour in considerable detail (see Boeing's granted patent US 7983996). Another five prioritised the self-preservation of the AI system; in other words, they were almost exclusively implementations of the Third Law.

The remainder, about half, were mainly concerned with the social and emotional behaviour of intelligent systems, that is, aiming to enhance their interaction with human beings and human society.

The following graphic shows the pattern in time of the applications filed, and of the number of patents granted.

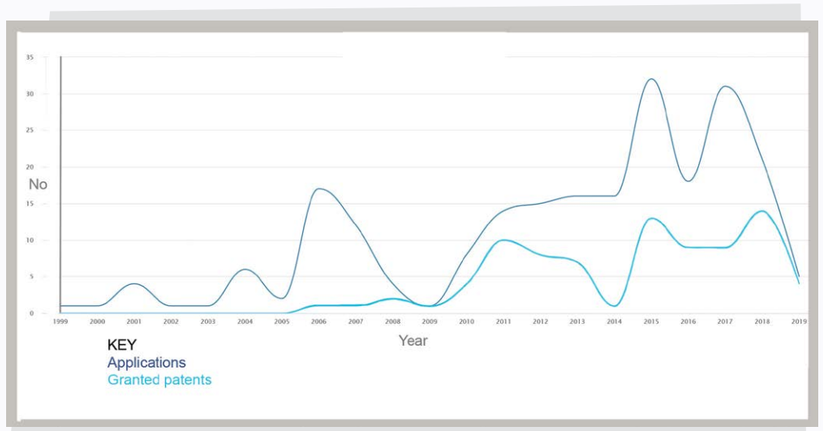

Fig. 2 Patent application and grant velocity for AI and ethics

This shows a rise in filings from 2006 to 2007, then a slump coinciding with the financial crisis, followed by a much higher filing rate from 2015.

Three cases explicitly invoked Asimov or the Three Laws of Robotics, and these tend to have very ambitious claims. Consider, for example, Alan Kadin's application US 2019258254 on a machine that constructs a virtual reality environment in which it can plan and implement goals:

> *… a conscious machine aware of its self, comprising: a set of sensor inputs from an environment; an artificial neural network, receiving the set of sensor inputs, and being configured to identify the self, agents, and objects represented in the set of sensor inputs, recognise correlated patterns in time and space in the environment, and plan achievement of a goal; and at least one automated processor, configured to construct a simplified dynamical predictive model of the environment that includes the self, interacting with the agents and the objects; a set of control outputs from the artificial neural network, dependent on the simplified dynamical predictive model, configured to alter the environment to implement the plan for achievement of the goal.*

Of the 11 cases to which only the First Law of Robotics seemed to apply - protecting a human - 10 were autonomously driven vehicles and one an emergency landing system for aircraft.

A further four cases, that appear to involve all three of the Laws of Robotics, build complex models of robot emotions as in Boeing's patent mentioned above, and contain modules which balance the requirements of the system against those of the human user, stated for example in Sony's US 2005197739 as:

> *… a behaviour control system and a behaviour control for the robot apparatus, having the function of adaptively switching between the behaviour selection standard taking the self-state into account, as required of the autonomous robot apparatus, and the behaviour selection standard taking the counterpart [human] state into account, depending on a prevailing situation.*

These often incorporate scenarios within which the AI system will act. US 2014288704 creates social robotic characters which may have a physical embodiment, that

> *…include behavioural provisions for some important but unlikely short-term interactive scenarios that could arise quickly, such as those involving emotional or practical support for users known to be at risk for distress, such as users diagnosed with dementia or autism.*

Conclusion

The growth in applications since 2015 has its origins in the deep learning revolution dating from about 2012 that resulted from the use of multilayer artificial neural networks to solve problems and outperform humans in fields such as beating the world champion in the game Go, visual pattern recognition e.g. for cancer diagnosis, and predicting protein folding and chemical targets for drugs. [2]

How will controls on robot behaviour such as Asimov's laws be brought to have their effect when the processes by which the robots arrive at their decisions are no longer detectable by humans as was the case in purely rule-governed AI systems? Perhaps we will just make a leap of faith and come to trust such "electronic persons" from our experience in interacting with them, as forecast in Steiner and Unbehaun's application US 2018012133.

100 years on from Asimov's birth and nearly 80 years since he first articulated the Three Laws of Robotics, we may finally be entering the future he created for us.

[1] Asimov, Isaac (1950). "Runaround". I, Robot (The Isaac Asimov Collection ed.). New York City: Doubleday. p. 40. ISBN 978-0-385-42304-5.

[2] "Why Deep Learning Is Suddenly Changing Your Life". Fortune. 2016. Retrieved December 1st 2019.