SAI Computing Conference 2016
July 13-15, 2016 | London, UK

Towards Ethical Robots: Revisiting Braitenberg's Vehicles

Christopher J. Headleand
School of Computer Science,
University of Lincoln,
Lincoln, UK,
cheadleand@lincoln.ac.uk

William Teahan
Llyr Ap Cenydd
Computer Science Dept.
Bangor University,
Bangor, UK,
{llyr.ap.cenydd, w.j.teahan}@bangor.ac.uk

***Abstract*—The development of software and machines capable of making ethical judgements is a topic of great interest with both the research communities and the public. Debates over the possibility and practicality of such systems have only intensified with the increased use of robotics in the military arena and the ubiquity of AI in commercial products. Modern innovations, such as the driverless car, will likely make artificial ethical agents a legal necessity. As a research field, it has received relatively little attention compared to other, more traditional, AI problems. In this paper, we propose a bottom-up reactive system that provides one possible solution. We will begin by describing the motivation to this work: the development of artificial ethical agents could both mitigate some fears about the future of autonomous AI, and providing insight into human moral reasoning. We then explore the related work, including the current attempts at simulating ethics. We describe our novel approach to ethical simulation, Vessels; a Braitenberg Vehicle inspired reactive agent approach. We, then, demonstrate how Vessels can be configured to simulate both Egoism and Altruism, comparing our simulations to the normative theory.**

Keywords—Simulated Ethics; Braitenberg Vehicles;

I. MOTIVATION

With the growing ubiquity of artificial intelligence, there is a greater need to develop systems that are not only smart, but ethical. This consideration is becoming increasingly important as robots and other artificial agents become more autonomous and away from human supervision. In many cases, machines are now in a position where they can impact the rights of humans [1].

However, other more consumer-level innovations such as the future of driverless cars is also fuelling the debate. For example, how should these vehicles behave in a crisis? What if an action to save a passenger resulted in the death of a pedestrian? For reasons such as this, the field of artificial ethical agents (AEAs) "is stepping out of science fiction and moving into the laboratory" [2].

There are many reasons why we may wish to build agents with the ability to act ethically. Providing autonomous entities with a concept of ethical behaviour may help avert dangerous or catastrophic consequences of rogue autonomous agents. For example, the ever increasing use of robotics and drones in war has lead to a significant amount of research on the ethics of so-called 'killer robots' [3], [4], [5], [6], [7], [8].

Similarly, some people have raised concerns about the long-term risks of AI research, and what danger it poses to humanity [9]. There is real concern within the public and some members of the research community that the development of artificial intelligence could be so hazardous that it would be incompatible with human life. These fears are not helped by science fiction horror stories; even the word 'robot' comes from a dystopian fantasy about a synthetic workforce who overthrow their human masters [10]. But if autonomous agents can be demonstrated that are capable of ethical reasoning, perhaps some of the risks involved in AI research can be avoided.

To look at the positive end of the debate there are a number of other reasons to research into ethical machines. For instance, if ethical judgement is a cognitive process, then maybe a superinteligent AI could make better ethical judgements than a human [11]? Furthermore, simulating ethical behaviour could provide insights into more human ethical reasoning judgements [12]. This could further the study of moral philosophy by providing theorists with tools to explore effects and consequences, substantiating their positions [13].

The majority of research into artificial ethical agents has explored the top-down approach. However, these machines are typically not robust enough to operate under real world conditions in real-time. Furthermore, they are limited by some of the fundamental problems in AI (such as the frame problem [14], [15]).

An alternative direction is the bottom-up approach [16]. The bottom-up approach has been successfully applied to a number of areas in AI, particularly in areas where the agent needs to operate in uncertain environments.

An example of these reactive agents is the *Vehicles* described by Braitenberg [17]. These uncomplicated agents are built by directly coupling sensors to actuators. However while relatively simple, these vehicles are able to exhibit seemingly complex behaviour by reacting to their environment. When describing their behaviour in psychology terms, they become interesting; their simple reactions can resemble complex human behaviour (such as love or aggression).

Braitenberg Vehicles are a major cornerstone of robotics research, and have been used as a platform to investigate a variate of behavioural phenomena. However, despite their use in the field of cognition, they have not been used as a tool to explore ethical behaviour. If we accept that the process of

978-1-4673-8460-5/16/$31.00 2016 IEEE

selecting ethical actions is a result of cognition, then could Braitenberg vehicles be used as the basis for an Artificial Ethical agent? In this paper we will explore this question.

For the purpose of this research, we will define an Artificial Ethical Agent as: *An autonomous entity that is capable of acting according to a set of morally defined considerations.*

II. Background

A. Simulating Ethics

The earliest attempts at building artificial systems capable of ethical judgement were decision support systems [2]. These efforts are especially prevalent in the medical arena such as systems designed to resolve biomedical ethical dilemmas [13]. Another example is ethical transplant software [18], capable of making superior judgements to humans [19]. There are also a number of practical models which have been proposed for the development of general ethical engines.

However, there have been relatively few attempts at building ethical reasoning into autonomous agents. This has left questions as to how these systems could be constructed.

Winfield et al. [20] describe one possible approach called a Consequence Engine. The Consequence Engine is a simulator embedded within an agent, allowing it to evaluate "what-if" scenarios. This allows it to experiment with a variety of possible actions before undertaking them in the real world. This approach is shown to work on simple situated robots in a limited world, with one robot changing its behaviour to prevent another from being harmed. However, the Consequence Engine approach is subject to the frame problem [14], [15], as it requires calculating the results of actions in a dynamic world. For this reason, its practicality is currently limited.

Bar-Cohen and Hanson [21] propose a similar idea as theoretical model for a general purpose ethical machine. Their model is based on a number of core requirements, (summarized by Sullins [22]):

1) estimate with high detail and accuracy the immediate state of the world around the agent;
2) predict the likely future states given the current possible candidate actions.

It could be argued that some current algorithms meet this criteria, such as MiniMax [23], but these only work in constrained worlds with a finite number of possible actions. Again, this model falls prey to the frame problem. It is also overly general, and could be used to describe almost any decision making process.

Another abstract model provided by Arkina [24] provides some alternative ideas and describes three ways an artificial ethical agent could be constructed.

Governor A system designed to intervene when an agent is about to conduct an action which would be considered unethical, a pseudo-conscience. It is named after the governors on steam engines, release valves which prevent the engine from running too hot.

Behaviour Controller A system that monitors all actions the agent undertakes to ensure they fall within a a set of ethical constraints.

Adapter A method of modifying either the *Governor* or *Behaviour Control* systems if an un-ethical action occurs despite intervention [25].

As with the previous two models, this is a top-down approach. While these models are logical on a theoretical level, they lack practicality within the current technological generation. Some researchers have gone as far as to question whether top-down approaches can be practically implemented [26]. For this reason, these approaches are mostly limited to toy problems in constrained worlds.

By contrast, we could construct artificial ethical agents in a bottom-up fashion, as described by Wallach et al. [16]. The bottom-up approach has been successful in many real-world problems, especially in the field of behaviour-based and reactive robotics.

A major advantage of reactive systems is that they operate in real-time, calculating the next, immediate action at each time-step. Due to this, they are ideally suited to unpredictable environments. Furthermore, the reactive approaches produce natural looking behaviour, often compared to the motion of insects.

This approach is arguably best illustrated by the *Vehicles* proposed by Braitenberg [17]. Braitenberg Vehicles are a thought experiment which demonstrate how seemingly intelligent actions can emerge from purely mechanical systems without the need for information processing. Through these simple devices, cognitive processes such as love and fear can simulated.

If Braitenberg Vehicles can be used to simulate cognition, it is reasonable to assume that we can use them to simulate ethics – making the assumption that ethics are a cognitive process. In the following subsection, we will describe Braitenberg Vehicles in order to introduce our own interpretation of this thought experiment.

B. Braitenberg Vehicles

In Valentino Braitenberg's seminal book 'Vehicles', a series of thought experiments are described which demonstrate how intelligent behaviour can emerge from sensorimotor coupling [17].

The core concepts of a Braitenberg Vehicle have been used in various studies of robotics and multi agent systems, both in physical and simulated experiments. For example they have been used to: simulate the movement of fish [27]; study group and herd behaviour [28], [29]; create teams of robots [30]; enable agents to localize the the source of a stimulus [31]; and explore the construction of abstract concepts [32]. Most of the examples of research in this area has focused on the types 2 and 3 vehicle, partly due to their simplicity but also because their implementation is well defined by Braitenberg.

Braitenberg's Vehicles are designed to autonomously navigate their environment through sensory input. Each sensor is directly coupled with an actuator (such as a motor), as when the sensor is activated, the actuator is driven. The amount it is driven is determined by the intensity of the stimulus. By changing how different sensors and actuators are configured on the vehicle, different behaviours can be produced, such as vehicles

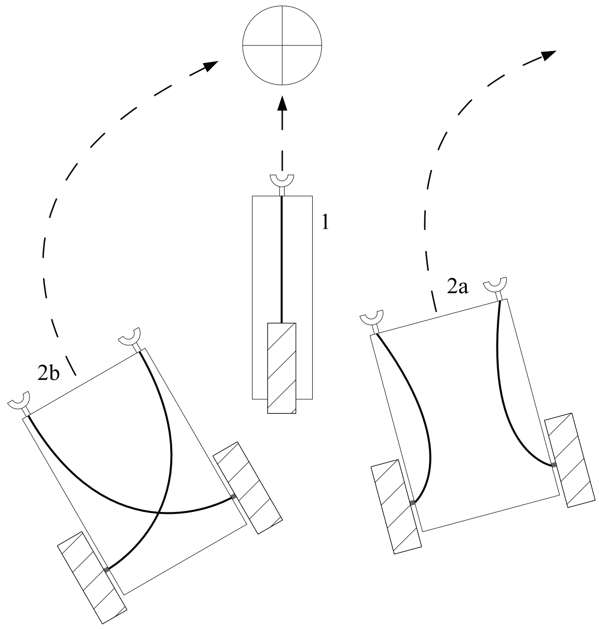

Fig. 1: Three different Braitenberg Vehicles in an environment with a stimulus (circle with cross). The more a sensor is stimulated, the more the connected actuator is driven. The type 2b (aggression) vehicle moves and turns towards the stimulus. The type 1 vehicle moves forward if the source is directly ahead of the sensor. The type 2a (fear) vehicle turns away from stimulus.

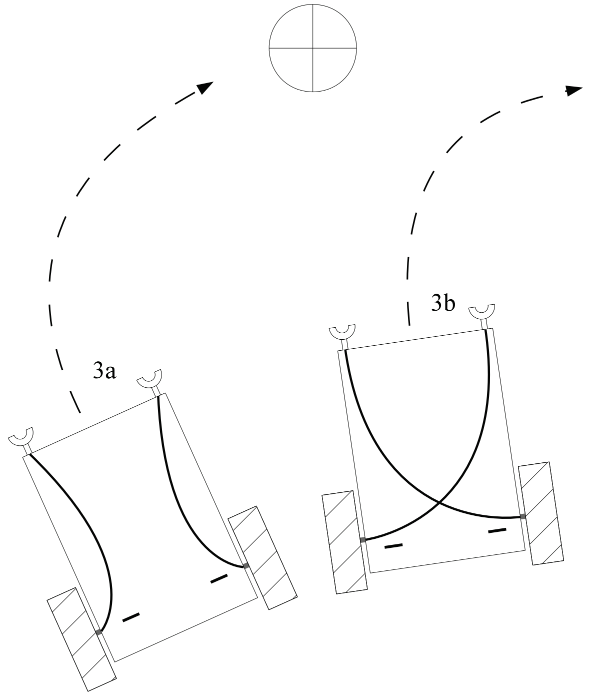

Fig. 2: The sensors on a class 3 vehicle inhibits the coupled actuators, slowing them down. A type 3a (love) vehicle will approach a light and slow when it gets close. A type 3b (explore) vehicle will steer away from a light and move.

that appear to be attracted to or repulsed by a light source. The simple Vehicles Braitenberg described were inspired by observations of animal brains. In his book, Braitenberg stated that the structures he observed were interpretable as pieces of machinery through their “simplicity or regularity”.

The vehicles’ hardware is too simple to be considered interesting from an engineering perspective. However, they are curious if their behaviour is described in psychological terms. The type 1 vehicle is described as being restless, because as soon as it is stimulated it moves. The 2a vehicle is described as being fearful of stimulus as it always turns away from it, and the type 2b vehicle is described as aggressive, as it will move with increasing speed towards a source(see Figure 1).

What makes this more interesting is that small changes in the vehicles’ connections produce significant changes in behaviour. Braitenberg said that type 2 vehicles were “crude” in their motion, as they are only ever excited by stimulus, and do not achieve restful states. In the type 3 vehicles, the connections between sensor and actuator were changed from positive to negative, inhibiting rather than exciting. This resulted in actuators that are driven when there is no stimulus, but slow as sensors are activated. A type 3a (love) vehicle moving towards a source will slow and stop when it gets close, rather than aggressively ram it (see Figure 2). This is a significant change in behaviour despite the relatively minor change in configuration.

Each vehicle type will exhibit increasingly complex behaviour as the complexity of the environment increases. For example, adding multiple sources of stimuli will have a dramatic effect on how the agent navigates the environment. Importantly, the behaviour is adaptive to the environment without the need for any information processing. To an observer it may appear intelligent, but the vehicle is simply reacting to its environment.

Vehicles have typically been used to explore behaviour by describing their actions using the language of psychology. Part of Braitenberg’s intention was to demonstrate how what appears to be complex behaviour can emerge from simple mechanisms interacting with an environment. This raises the question whether the same philosophy can be applied to the study of ethics. Could we use the language of morality to describe the actions of the Vehicles?

While imbuing these simple devices with ethical reasoning could be criticized, we would stress that it falls within the original intentions of Braitenberg. We would also argue that describing a vehicle using an ethic (such as Egoism) is equivalent to describing them using an emotion (such as love).

III. Ethical Vessels

In this section, we will introduce Vessels, a variation on Braitenberg Vehicles designed to explore the problem of simulating ethics. We have named them *Vessels* to avoid confusion with *Vehicles*.

Following the philosophical approach of Braitenberg, we will describe our Vessels as if they were creatures in a natural environment. Furthermore, we will describe their behaviour using the same ethical vocabulary that we would for natural creatures. With this considered, it is important to establish that we are not trying to claim that Vessels represent how natural creatures make ethical judgements. We are simply using them as a device to explore the problem of simulating ethics.

The purpose of Vessels is to demonstrate that behaviour resembling ethical phenomena can emerge, through an agent's interaction with the environment and other agents. By altering the agent's internal couplings, we can create various different classes of ethic-like behaviour. This is without the need for information processing and complex top-down systems, such as the Consequence Engine described in the background to this work.

While some may argue that the bottom up approach is weak AI, our position is that there is little benefit in limiting any agent to the natural mode of decision making. Furthermore, some researchers (notably Brooks [33]) have argued that reactive, bottom-up approaches could be a model for all natural intelligence. However, it is not our intention to contribute to this debate. We would simply propose that for the purpose of simulating ethics, if an action appears to be motivated by moral considerations, we should treat the action as being an ethical one.

Our position, working with observation, is consistent with the study of ethics. Coeckelbergh [34] notes that the study of human morality relies on observational insights. For consistency, we should apply the same assessment standards to other entities. This standpoint, that could be described as anthropomorphic, opens up new opportunities into gaining insight into human ethics through simulation.

In the following sections, we will provide an argument that a Vessel is an example of an ethical agent as we have interpreted the term. We also argue that the Vessels are examples of implicit Ethical Agents, based on the taxonomy provided by Moor [35].

IV. The Two Lights Experiment

The *Two Lights Experiment* is a simple environment, devised to investigate the most rudimentary behaviour that could be described as ethical. It is named after two types of light found in the world. The first type is a light situated in the environment as a stimulus (referred to as the resource). The second type is a light mounted on the top of each Vessel (referred to as the valence) used as a status indicator.

Each Vessel is powered by light, utilising light as a resource to charge their batteries. Their solar cells only activate under the intense white light generated by resources. This restriction means that they must get close to a resource to charge, otherwise their power slowly drains.

The valence light on each Vessel provides an indication of the current state of that agent. If they are content, the valence glows blue; when in pain, the light glows red. Taken as a spectrum (red-blue), the valence indicates the sum of each Vessels' current welfare.

The closer the Vessel is to a resource, the higher its pleasure will be as it is in an ideal position to charge its batteries. We can refer to this abstractly as *eating*. As a corollary, the Vessel will feel pain as the power in its batteries is depleted, which we can refer to as*hunger*.

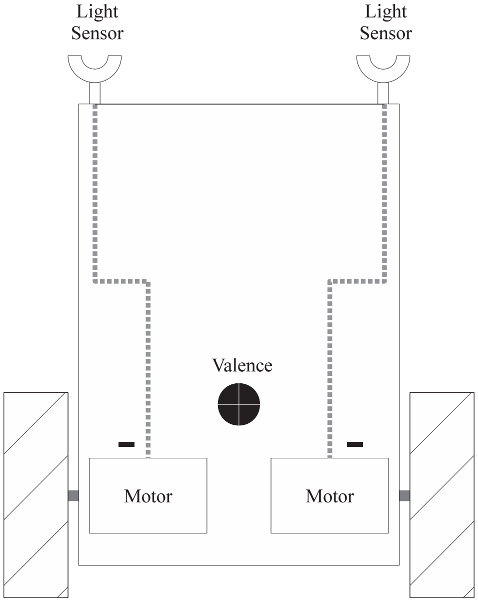

Fig. 3: A diagram of a basic Ethical Vessel in the style of a Braitenberg Vehicle. At the front of the vehicle, two light sensors can be seen, while at the back two motors independently drive two wheels. The black circle with a white cross symbolizes the valence light which changes colour to reflect the current welfare of the Vessel. For the sake of simplicity, the battery and solar cells have not been included on the diagram.

To determine the colour of the valence light, we use the following formula:

$$V = (\frac{(d-r)}{(R-r)}) + (p-1).$$

In this formula; V is the welfare/valence value, scaled to the range -1-$+1$. R is the maximum distance from the resource a Vessel can be whilst still charging, r is the radius (from the resource) inside which no additional benefit is obtained (maximum intensity). d is the Vessel's actual distance from the resource (visualized in Figure 4). Finally, p is the current charge of the battery (between 0 and 1).

The welfare value (V) is mapped to the valence light on the back of the Vessel. $+1$ would result in pure blue, -1 would result in pure red, indicating the agent's current welfare state.

V. Simulating Ethics

In the following two sections, we will describe two Vessels, one designed to exhibit Egoism-like behaviour, the other to act in an altruistic fashion. Both subsections will follow a similar format; we will introduce a normative ethical theory, then describe a Vessel model which could (in practice) produce similar behavioural phenomena. We conclude the section by placing the Vessel within the Two Lights experiment to observe its behaviour.

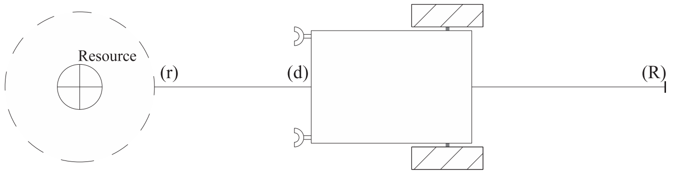

Fig. 4: A Vessel approaching a resource. (r) is the range at which the light can be used at maximum efficiency. (R) is the maximum range the Vessel could be from the resource and still charge. (d) is the current distance of the Vessel from the resource.

All simulations have been conducted with the same basic parameters. 15 Vessels of the prescribed type were placed in an environment with either 1, 2 or 3 resources. Once the simulation was initialised, the agents moved based on their sensorimotor couplings. The environment was toroidal, wrapping around at the 4 boundaries. At the end of 10,000 ticks, the number of surviving agents were counted. Each simulation configuration (1, 2 and 3 resources) was repeated 100 times using the NetLogo simulation language/engine.

In each simulation, the type of Vessel is highlighted by a change of marker colour in figures—blue for Egoism, yellow for Altruism. If a Vessel depleted its battery, it is depicted by a grey marker. A Vessel that has run out of battery will be described as 'dead'.

Although we include the background to each normative ethical theory below, we are not suggesting that the Vessel is recreating that behaviour precisely. We simply note that phenomena, which could be described as ethical, can emerge from the Vessel models. Furthermore, it is important to not focus on the specifics of the design. To paraphrase Braitenberg, the realization of the idea is less important than the idea itself. While we attempt to describe something that may be fabricated, we have also simplified the design to aid discussion of each Vessel type. Should someone wish to build physical versions of a Vessel, technical issues would need to be overcome (which we will describe in the relevant sections). As we are simulating the Vessels, this is not a limiting factor of this work.

VI. Egoism

The normative theory of Egoism holds that the needs of the individual should be the motivation of one's actions above all others [36], [37]. Within egoism there is no moral obligation to serve or protect the needs of others. It has been argued that Egoism is the only ethical position which reflects the rights of the individual. Some supporters of this position have also argued that Altruism (the opposite position) is destructive, as an individual views their life only as something to be sacrificed for the good of others [38].

A. Vessel Model

To simulate Egoism with Vessels, we need only the individual agent. There is no obligation to consider the needs of others. In the two lights experiment, the agent only has one need, to get close enough to the resource to charge and survive. This can be achieved simply by recreating the circuity from the type 3a (love) Braitenberg Vehicle. This configuration (inhibiting motors) causes the agent to steer towards, and then stop close to a source of stimulus within the environment. As Egoism is essentially a selfish theory, we do not need to consider any other connections to sense the other agents in the environment. However, it is possible that a Vessel's movements could block others from reaching the resource.

B. Simulation Results

Where resources are plentiful (our three resource environment), this tactic appears to work reasonably well. The lack of competition means that in the majority of cases, the Vessel will reach a resource unobstructed by other agents. However, as the number of resources in the environment is decreased, the chance of survival is lower. This is demonstrated in figure 5:(a) and (b) – note that in the 2 light environment, more Vessels have 'died'.

In a single light environment (shown in figure 5:(c)), even fewer Vessels survived throughout the experiment. There was not enough space around the light to sustain all the agents within a useful radius. As the Egoist Vessel is entirely motivated by self interest, it does not move to allow others to charge, causing the blocked agents to die. This is the main limitation of the Egoist Vessel, a lack of cooperation causing an undue amount of the population to perish.

VII. Altruism

An Altruist believes that individuals have an obligation to help others without concern for their own needs. Altruistic behaviour is generally defined as a costly act that confers benefits on other individuals [39]. As a normative position, it is generally considered to be the direct contrast to Egoism.

A. Vessel Model

To design the altruistic Vessel, we will build upon the Egoism model. This already achieves the basic objective of producing a Vessel that will seek out the resources in its environment. The valence telegraphs the agent's own welfare. However, the Vessel is unable to recognize the welfare of others. To add this capability, we can add a colour sensor to the Vessel (such as the colour sensors described by [32]). We will refer to this as the empathy sensor, as it will allow the agent to recognize the welfare of other agents by detecting ambient light along the red-blue spectra.

This sensor is placed above the agent's own valence light, with a shield between the two preventing it cross-detection. [1] The result is that it only considers the welfare of other agents.

We can consider the empathy as being two independent streams of information, one that just detects red light, the other detecting blue light (we will refer to these as colour channels). Both colour channels are directly coupled to the Vessel's motors. If the blue colour channel is stimulated, the

[1] If we were trying to build a physical Vessel, then this approach would probably not work. Simply put, it would be impossible to completely shield the agent's empathy sensor from its valence. However, we have used lights purely for descriptive purposes. Should we want to fabricate these agents, other hardware (such as bluetooth) could be used.

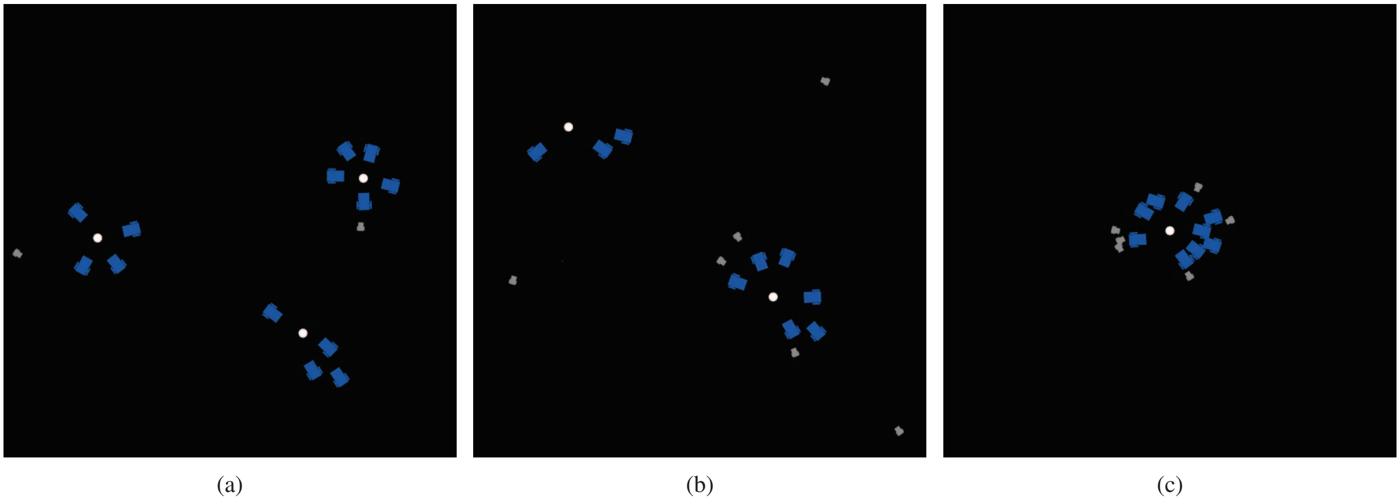

Fig. 5: If there is little competition for resources in the environment (a) then Egoism is a suitable strategy, and the majority of agents will survive. However, as resources become more limited (b), competition for the resources will result in less agents surviving. In environments where resources are even more scarce (c), there is a much higher probability that the Egoist Vessel will not survive. This is because of a higher risk that one Vessel will block another from reaching the resource, causing it to lose power and die.

motors will be inhibited, slowing them down, in proportion to the intensity. By contrast, if the red channel is stimulated, the motors will be driven, causing the Vessel to accelerate.

If the Vessel is surrounded by other Vessels with positive welfare (blue light), it will be in a restful state, its motors fully inhibited. However, if those around it are experiencing negative welfare (red light), it will move. Consider a scenario where one Vessel is blocking another from a resource. The blocked Vessel would eventually go into negative welfare and its valence would glow red. The blocking Vessel would detect this, through its empathy sensor, driving its motors to move it out of the way. Eventually a stable state results, the best result for all remaining Vessels.

However, there is one limitation with this system: the agent is currently more Utilitarian than Altruistic. It considers the aggregated welfare of all local Vessels. For example, if the majority of local agents around a Vessel had positive welfare, the agent may seemingly ignore the one that is dying. We can compensate for this by limiting the influence of individual color channels running them through a simulated variable resistor. For our experiments, the blue channel was given a bias of 0.2. This causes the Vessels to consider their negative valence neighbours before those that are content. If we were to describe this using the language of ethics, we could refer to is as a *charitable* Vessel.

B. Simulation Results

When simulated, the altruistic Vessel is restless. If it is currently occupying a resource, whenever another Vessel is nearby with a low welfare, it will move on, often without fully charged batteries. This restlessness results in a large number of Vessels dying before the end of the simulation. This is not always logical, or the best for the group. Our simulation seems to support this position, despite the simplicity of our Vessels.

Table I highlights the average survival rates for the two Vessel types in each of the three environment configurations (1, 2 and 3 resources). Firstly, we can see that the chance of survival is proportional to the number of resources in the environment. Secondly, we can see that the Altruist Vessel is significantly less likely to survive, although survival rates on both Vessel types are still rather low.

The Altruist Vessels have the opposite problem of the Egoist Vessels. In the Egoism experiments, a Vessel was most likely to die because it was blocked by another agent. An Altruist Vessel was most likely to die because it moved away from a resource to allow a blocked agent in. If a balance could be created between these two positions, then the strengths of one could possibly alleviate the weaknesses of the other.

VIII. Value Systems

While our model systems are interesting from a research perspective, it could be argued the practical use is limited. The fact is, humans are rarely entirely egotistical or altruistic. We vary based on our own moral codes and the current situation we find ourselves in. It stands to reason that machines may also need to vary their ethical response.

TABLE I: Survival rate data from the Egoism and Altruism simulations. Each Vessel was simulated 100 times for each environment configuration (1, 2 and 3 resources). Every simulation started with 15 agents.

	Environment Setup (# Resources)		
Vessel Type	3	2	1
Egoism	11.2	9.5	7.9
Altruism	9.1	6.7	5.2

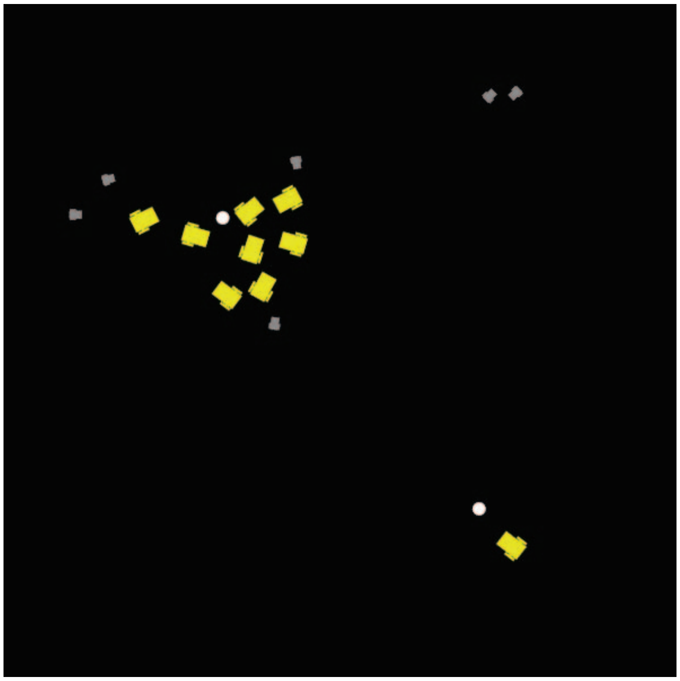

Fig. 6: In simulations, the Altruist Vessel would sacrifice itself in the event of blocking another from the resource. Unfortunately, this resulted in a large number of Vessels not surviving the duration of the simulation.

One way this could be approached is to adapt our Vessels with a value system. A value system is a set of consistent ethical responses. These systems generally contain exceptions which both rank potential courses to resolve contradictions. A common example of a synthetic value system is the ‘Three Laws of Robotics’ described by Issac Asimov in Runaround [40].

The Three Laws are stated as (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. (2) A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law. (3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law [41]. The exceptions in the system allow the robot to resolve conflicts, in principle, between the individual laws.

In our motivation, we discussed the ethical choices of driverless vehicles. Specifically, the problem of whom the vehicle should protect in the event of a imminent crash. Should a vehicle kill the passengers, if it meant saving pedestrians? If we decided that the vehicle should not allow its passengers to die, we have the basis of a simple value system.

1) The Vehicle must avoid impacts with pedestrians.
2) The Vehicle must protect its passengers, except where such protection would conflict with the first value.

This allows the ethical judgement of the vehicle to change based on the current situation expanding the range of events that can be handled, and resolving contradictions. If the same philosophy is applied to the Vessels, we could design a value system that causes a Vessel to act in its own self interest, but also be altruistic when required. We could describe this type of behaviour as showing compassion.

A. Threshold Based Value System

We have discussed the results of the true Egoist and true Altruist Vessels when placed in the two lights environment. Neither experiment resulted in all Vessels surviving. Our goal is to have as many Vessels survive as possible, ideally the entire population. We noted that in the egoist experiment, Vessels died because others blocked them from the light source. However, in the altruist experiment the opposite was true; Vessels died because they constantly moved to allow others access.

We propose that it is possible for a Vessel to embody both these normative positions simultaneously, using a value system. The value system would allow the Vessel to be selfish when its own life was at risk (Egoist), and selflessly move aside when those around it are in need and its welfare is not at great risk.

A candidate value system will be described with two rules:

1) A Vessel must act to preserve its own existence.
2) A Vessel must not prevent another from self preservation, except in situations where any sacrifice would conflict with the first rule.

B. Vessel Model

To implement our value system on an ethical Vessel, we begin with the Altruist model. As this is built upon the Egoism model, we only need a method to switch off the altruistic functions when the Vessel is at risk, essentially devolving the Vessel to a simpler, selfish state whenever it must protect its own interests.

We can build the Compassion Vessel by adding a circuit (called the value trigger) to the wires leading to the Vessel’s valance light. The value trigger is connected to a relay on the Vessels empathy sensor. The trigger itself is activated whenever the welfare of the vehicle is positive (blue valence). If the value trigger is active, then the relay is held open, allowing the Vessel to be influenced by the others around it. When the trigger is not active, the relay closes, switching off the empathy sensor and the altruistic behaviour.

When the agent has positive welfare, and not currently at risk, it will act like the original altruistic Vessel. It will continue to move while other Vessels are expressing negative valance. This satisfies our second rule of the value system, that the Vessel should not prevent other Vessels from self preservation. However, once the Vessel’s own welfare drops below 0 (into red valence) its empathy sensor is cut off, causing the Vessel to act only in its own interest (becoming an Egoist).

When simulated, the Compassion Vessel exhibits emergent behaviour. Initially, all Vessels will attempt to steer towards the light source. As soon as a Vessel arrives, it will wait there charging, following the Egoist behaviour. As expected, this behaviour results in some Vessels being blocked from reaching the light source. Once a blocking Vessel’s welfare becomes positive, that Vessel will move allowing another access to the resource. This forms a continuous cycle allowing all the

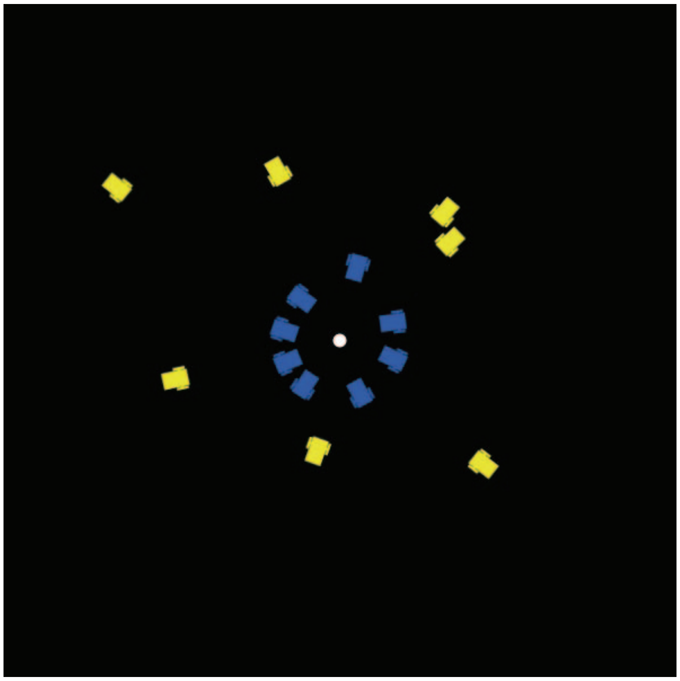

Fig. 7: The Compassion Vessel in simulation. This figure depicts eight (blue) vessels circled around a resource in the centre of the environment. These vessels are currently observing the first rule of their value system, and protecting their own existence by acting as egoists. Seven yellow vessels have recently moved away from the resource allowing the blue vessels to access, acting altruistic, and following the second rule of their value system.

Vessels (in the vast majority of cases) to survive the test. In the one light environment, an average of 14.6 Vessels, over the 100 simulations, survived the full duration of the experiment.

IX. Conclusion

In this paper, we have argued that the reactive approach provides a possible route to the simulation of ethical behaviour and the production of artificial ethical agents. For this work, we have focused on using Braitenberg Vehicles. We proposed a re-interpretation of Braitenberg's work, which we have termed "Ethical Vessels". We demonstrated how the core concepts of two normative theories could be simulated, highlighting the flaws of any single normative approach. We further developed our theory, demonstrating how a simple value system could be embedded into these Vessels, highlighting the significant improvements this technique provides.

While the current experiments utilize simple agents, it is our hope that these can be used as a platform for more complex ethical reasoning. Just as Braitenberg Vehicles have been used in the field of cognitive science and behaviour based robotics, we hope that the Ethical Vessel can provide similar insights and developments.

Although there is a significant conceptual gap between what we have produced and something which could be described as human-like, we believe this is ultimately possible within a reactive system. Vehicles (and by extension Vessels) are modelled on insect intelligence, and we do not see this as a long-term limitation. This is because it is a relatively small evolutionary jump between insects and humans [42].

A. Future Work

This work can be developed to simulate additional normative behaviours. We also intend to look at more complex ways to blend behaviours rather than a simple, threshold-based value system. Following this, we intend to use Vessel-based ethical reasoning in simulated driverless cars to evaluate their behaviour during collision events. In this area, a Vessel has a major advantage over other comparative systems. Its reactive design allows it to operate in real-time, which is essential during high-speed collisions.

Acknowledgement

The authors extend their thanks to HPC Wales for their ongoing support. Additionally, Christopher Headleand thanks Fujitsu for their ongoing financial support of his PhD and research activities.

References

[1] B. C. Stahl, "Information, ethics, and computers: The problem of autonomous moral agents," *Minds and Machines*, vol. 14, no. 1, pp. 67–83, 2004.

[2] W. Wallach, "Implementing moral decision making faculties in computers and robots," *Ai & Society*, vol. 22, no. 4, pp. 463–475, 2008.

[3] M. Coeckelbergh, "Drones, information technology, and distance: mapping the moral epistemology of remote fighting," *Ethics and information technology*, vol. 15, no. 2, pp. 87–98, 2013.

[4] T. Hellström, "On the moral responsibility of military robots," *Ethics and information technology*, vol. 15, no. 2, pp. 99–107, 2013.

[5] A. Krishnan, *Killer robots: legality and ethicality of autonomous weapons*. Ashgate Publishing, Ltd., 2009.

[6] P. Lichocki, A. Billard, and P. H. Kahn, "The ethical landscape of robotics," *Robotics & Automation Magazine, IEEE*, vol. 18, no. 1, pp. 39–50, 2011.

[7] J. McMahan and B. J. Strawser, *Killing by remote control: the ethics of an unmanned military*. Oxford University Press, 2013.

[8] R. Tonkens, "Should autonomous robots be pacifists?" *Ethics and information technology*, vol. 15, no. 2, pp. 109–123, 2013.

[9] F. de Waal, "The animal roots of human morality," *New Scientist*, vol. 192, no. 2573, pp. 60–61, 2006.

[10] K. Čapek, *RUR (Rossum's Universal Robots)*. Penguin, 2004.

[11] N. Bostrom, "Ethical issues in advanced artificial intelligence," *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–286, 2003.

[12] P. Danielson, "Designing a machine to learn about the ethics of robotics: the n-reasons platform," *Ethics and information technology*, vol. 12, no. 3, pp. 251–261, 2010.

[13] M. Anderson, S. L. Anderson, and C. Armen, "An approach to computing ethics," *Intelligent Systems, IEEE*, vol. 21, no. 4, pp. 56–63, 2006.

[14] D. Dennett, "The frame problem of ai," *Philosophy of Psychology: Contemporary Readings*, vol. 433, 2006.

[15] M. Shanahan, "Frame problem, the," *Encyclopedia of Cognitive Science*, 2006.

[16] W. Wallach, C. Allen, and I. Smit, "Machine morality: Bottom-up and top-down approaches for modeling human moral faculties," *AI & Society*, vol. 22, no. 4, pp. 565–582, 2008.

[17] V. Braitenberg, *Vehicles*. Cambridge, MA: MIT Press, 1984.

[18] F. A. Hanson, "Beyond the skin bag: on the moral responsibility of extended agencies," *Ethics and information technology*, vol. 11, no. 1, pp. 91–99, 2009.

[19] E. de Sevin and D. Thalmann, "A motivational model of action selection for virtual humans," in *Computer Graphics International*. IEEE, 2005, pp. 213–220.

[20] A. F. Winfield, C. Blum, and W. Liu, "Towards an ethical robot: internal models, consequences and ethical action selection," in *Advances in Autonomous Robotics Systems*. Springer, 2014, pp. 85–96.

[21] Y. Bar-Cohen and D. Hanson, *The coming robot revolution: Expectations and fears about emerging intelligent, humanlike machines*. Springer Science & Business Media, 2009.

[22] J. P. Sullins, "Robowarfare: can robots be more ethical than humans on the battlefield?" *Ethics and Information technology*, vol. 12, no. 3, pp. 263–275, 2010.

[23] S. Russell, P. Norvig, and A. Intelligence, "Artificial intelligence: A modern approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, 1995.

[24] R. C. Arkina, "Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture," in *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference*, vol. 171. IOS Press, 2008, p. 51.

[25] ——, *Governing lethal behavior in autonomous robots*. Taylor and Francis, 2009.

[26] W. Wallach, "Robot minds and human ethics: the need for a comprehensive model of moral decision making," *Ethics and Information Technology*, vol. 12, no. 3, pp. 243–250, 2010.

[27] X. Tu and D. Terzopoulos, "Artificial fishes: Physics, locomotion, perception, behavior," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*. ACM, 1994, pp. 43–50.

[28] R. S. Olson, A. Hintze, F. C. Dyer, D. B. Knoester, and C. Adami, "Predator confusion is sufficient to evolve swarming behaviour," *Journal of The Royal Society Interface*, vol. 10, no. 85, p. 20130305, 2013.

[29] R. S. Olson, D. B. Knoester, and C. Adami, "Critical interplay between density-dependent predation and evolution of the selfish herd," in *Proceedings of the 15th annual conference on Genetic and evolutionary computation*. ACM, 2013, pp. 247–254.

[30] C. R. Kube and H. Zhang, "Collective robotics: From social insects to robots," *Adaptive behavior*, vol. 2, no. 2, pp. 189–218, 1993.

[31] A. Lilienthal and T. Duckett, "Experimental analysis of smelling braitenberg vehicles," *environment*, vol. 5, p. 10, 2003.

[32] C. Headleand, L. Ap Cynedd, and W. J. Teahan, "Berry eaters: Learning colour concepts with template based evolution evaluation," in *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, vol. 14, pp. 473–480.

[33] R. A. Brooks and J. H. Connell, "Asynchronous distributed control system for a mobile robot," in *Cambridge Symposium_Intelligent Robotics Systems*. International Society for Optics and Photonics, 1987, pp. 77–84.

[34] M. Coeckelbergh, "Moral appearances: emotions, robots, and human morality," *Ethics and Information Technology*, vol. 12, no. 3, pp. 235–241, 2010.

[35] J. H. Moor, "The nature, importance, and difficulty of machine ethics," *Intelligent Systems, IEEE*, vol. 21, no. 4, pp. 18–21, 2006.

[36] P. A. Facione, D. Scherer, and T. Attig, *Values and society: An introduction to ethics and social philosophy*. Prentice Hall, 1978.

[37] J. Rachels, "Ethical egoism," *Ethical Theory: An Anthology*, vol. 14, p. 193, 2012.

[38] A. Rand, *The virtue of selfishness*. Penguin, 1964.

[39] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, no. 6960, pp. 785–791, 2003.

[40] I. Asimov, "Runaround," *Astounding Science Fiction*, vol. 29, no. 1, pp. 94–103, 1942.

[41] L. McCauley, "Ai armageddon and the three laws of robotics," *Ethics and Information Technology*, vol. 9, no. 2, pp. 153–164, 2007.

[42] R. A. Brooks, "Intelligence without representation," *Artificial intelligence*, vol. 47, no. 1, pp. 139–159, 1991.