https://orcid.org/0000-0002-5509-4185
https://orcid.org/0000-0002-7639-0119
https://orcid.org/0000-0002-4309-6511

UDC 004.832.2

J. PAGE[*], F. MUKHLISH[*, **, *], M. BAIN[****]**

ENSURING THE ALIGNMENT OF GENETIC/EPIGENETIC DESIGNED SWARMS

[*]School of Mechanical and Manufacturing Engineering, University of New South Wales, Sydney, Australia
[**]Engineering Physics Research Group, Institut Teknologi Bandung, Bandung, Indonesia
[***]Center of Instrumentation Technology and Automation, Institut Teknologi Bandung, Bandung, Indonesia
[****]School of Computer Science and Engineering, University of New South Wales, Sydney, Australia

Анотація. *Однією з головних задач дослідників і розробників штучного інтелекту (ШІ) є забезпечення відповідності систем прагненням людей, які з ними взаємодіють. Ця проблема стає ще складнішою для систем, для яких розробляються власні правила їх функціонування і в яких задіяні кілька агентів. У статті розглядається ряд наслідків використання методів генетичного/епігенетичного проєктування за умов, коли структура контролю розробляється без безпосередньої участі людини. Це створює особливі труднощі у забезпеченні відповідності протоколів контролю бажанням підбурювачів і уникнення завдання непередбачуваної шкоди. У роботі також досліджуються випадки ще більшого ускладнення задачі, коли система ШІ має багато агентів. Сучасні системи керування часто є децентралізованими, що забезпечує більш надійне рішення, ніж при використанні центрального контролера. Конкретним прикладом реалізації цього підходу є групи, які самоорганізовуються, і в яких агенти діють незалежно від центрального контролю. З точки зору вирівнювання, це породжує ряд певних проблем. В інтересах людини повинні діяти не тільки окремі агенти, але й група як колектив. Для однорідної групи це досить складно, але пропозицій щодо використання гетерогенної поки що не було. Були і наразі продовжуються численні дослідження та дискусії стосовно того, як створити глобальну етику ШІ та яку форму вона може прийняти, але прогрес насправді є дуже повільним. Частково це пояснюється тим, що навіть Загальна декларація прав людини має ряд недоліків. Усі країни, які підписали цю Декларацію ООН, вважають, що вони принаймні намагаються її реалізувати. Проте проблема полягає в її тлумаченні, адже багато підписантів вважають, що інші порушують її положення. Те ж саме стосується і будь-якої загальної угоди про етику ШІ. У цій статті пропонується рішення, в якому базова етика систем ШІ хоча і є індивідуальною, проте має відповідати вимогам у випадках взаємодії з іншими утвореннями ШІ або людьми.*

Ключові слова: *генетичні/епігенетичні алгоритми, вирівнювання ШІ, етика ШІ.*

Abstract. *One of the major concerns of AI researchers and implementers is how to ensure that the systems stay aligned with the aspirations of the humans they interact with. This problem becomes even more complex for systems that develop their own operational rules and where multiple agents are involved. The paper addresses some of the implications of using genetic/epigenetic design techniques where the control structure is developed without direct human involvement. This presents particular difficulties in ensuring that the control protocols stay aligned with the desires of the instigators and do not cause unpredicted harm. It also explores how this problem is further complicated when the AI system has many agents. Modern control systems are often decentralized which provides a more robust solution than using a central controller. A specific example of this approach is Self-Organising Swarms where the agents act independently of the central control. From an alignment point of view, it generates particular problems. Not only must the individual agents act in the best human interest but the swarm as a collective must do it as well. This is difficult for a homogeneous swarm and no proposal for a heterogeneous one has yet been made. There have been and continue to be considerable research and discussions on how to create and what form a global AI ethics might take, but any progress has been slow. This is partly because even the*

Universal Declaration of Human Rights has difficulties. All the nations that have signed up to the UN Human Rights Declaration believe they are at least trying to implement it. The problem is in the interpretation where many signatories believe others are in breach. The same would apply to any universal AI ethics agreement. This paper proposes a solution where the AI systems' basic ethics are individual but have to comply where they interface with either other AI entities or humans.

Keywords: *genetic/epigenetic algorithms, AI alignment, AI ethics.*

DOI: 10.34121/1028-9763-2022-1-3-11

1. Introduction

Many researchers and scholars have expressed concern about how to ensure Artificial Intelligent systems act in the best interests of humanity, often referred to as alignment. For example, the late Stephen Hawking claimed on the BBC, "The development of full artificial intelligence could spell the end of the human race." [1] Views of this nature, whether well-founded or not, could lead to restrictions being placed on AI research and implementation and even some AI systems being banned. The concerns expressed by Hawking and others have led to the establishment of a number of research organizations such as The Centre for the Study of Existential Risk [2] and Future of Life Institute [3]. Most of these researchers, however, concentrate on the risks of Artificial General Intelligence and beyond which has yet to be proven feasible. In fact, even Low-Level Artificial Intelligence which is already prevalent in our society presents significant risks [4]. While they may not present an existential risk, the collapse of utilities such as power or water systems could lead to significant loss of life and social disruption.

Many computer-based systems rely on distributed logic as it delivers significant advantages in robustness and this is the basis for swarm robotics. The problems with such systems are that the control programming becomes extremely complex very quickly, so one has to utilize automatic programming. In our case, we use Genetic algorithms modified by overlaying Epigenetic algorithms. This, however, leads to a surrendering of direct control of the process, making it hard to ensure alignment. For the "end", this is manageable, but for the "means" which also have to be aligned, this is much more difficult. "The end justifies the means" is not acceptable [5, 6].

The aim of this paper is to address some of the issues that will, however, require more profound research and discussion before a practical solution is achieved. In the paper, we look at some of the particular concerns associated with autonomous control development using genetic/epigenetic self-learning systems. It also addresses the advantages of a local AI ethics over a global one.

2. Control system for a self-organizing swarm

Self-organizing swarms operate with very little or no central control. As such, they have to structure their own tactics for completing a mission based on a set of rules. Though Reynolds [7] demonstrated, a swarm could be made to function with only three simple rules in practice to carry out a significant task, many more rules are required. As the task becomes more complex so the number of rules tends to grow dramatically as is demonstrated when exploring granular computing in rule-based systems [8]. A complication is further added to if the environment is dynamic. The classic method for addressing this problem is to use a behaviour-based method. It consists of developing an initial model, its implementation, evaluation, and modification. This process can continue until an acceptable solution is reached. The problem is that as the model gets more and more complex, it becomes difficult to generate and understand it. Three methods are often used to address this problem: Probabilistic Finite State Machine (PFSM), virtual physics-based design, and stigmergy, but they all have limitations as the complexity grows and there are problems with dynamic environments. An alternative approach is to allow the system to evolve to meet the requirements. One way of achieving this, though it still has significant limitations, is to use genetic algorithms. This can, however, be improved by wrapping an epigenetic algorithm around the genetic algorithm.

3. A genetic algorithm approach to improving behaviour

An evolutionary computational [9] approach for control system design is based on encoded characteristics of the control strategies being grouped into artificial chromosomes [10]. Each chromosome represents some particular characteristic for each strategy. Its fitness value is evaluated based on a given fitness function. The chromosomes (strategies) with high fitness value are allowed to breed through the normal genetic operators of recombination, mutation, and selection. Progenies with higher fitness ratings will replace the current ones with a lower fitness rating in the population. The process is repeated until the fitness value of the new generation meets the designated criterion. This method of generating control strategies in robotics is defined as Evolutionary Robotics (ER) (see Fig. 1) [11].

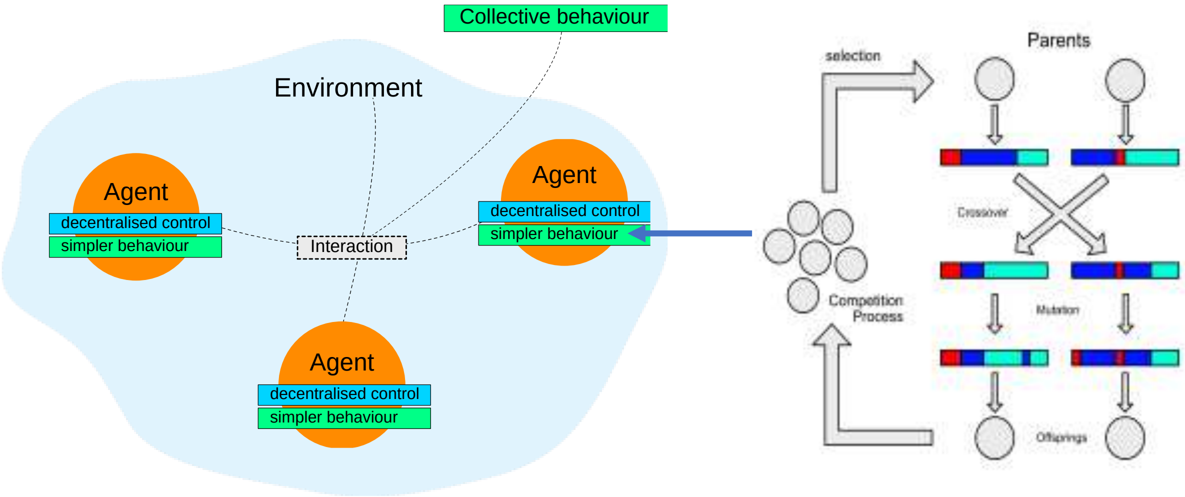

Figure 1 – Evolutionary Swarm Robotics from Nolfi et al., 2016

Despite the strength of this approach, it does present some fundamental issues. In Pure Darwinian Evolution [12], fitness is almost completely driven by survival, with limited sexual selection. So, for example, if a species lives on a volcanic island, some will evolve to take advantage of foraging in the shallow water, while some will develop the ability to climb and exploit the volcanic slopes. If the sea level changes so that it rises until the island is submerged, those adapted to mountain foraging will die, and the reverse is true if the sea recedes to the deep ocean. If the process is slow, then they may adapt it, but this is not based on constant adaption to the changing environment and has no element of prediction. Though Lamarck tried to address this problem, it was only with the introduction of the concept of Epigenetics that a workable explanation was found.

3.1. Deception

This is mainly due to the difficulty of determining a fitness function (objective function). In order to determine whether an evolved chromosome is better than its parents a measure is required. This has to be defined, and the selection of such a measure is not simple. Ideally, it is towards this target the control system is evolving, so it must be a well-defined finish point that meets all the requirements of the system. It is also important that it does not cause the system to migrate to a local solution. This is further complicated if the system that is being controlled is dynamic as the desired objective function may also change.

3.2. Exploration and exploitation dilemma

This is a problem for all data search engines and is no less a problem for genetic algorithms. In the present context, the problem is how long to continue searching for a better solution and when to concentrate on improving the existing solution. To some extent, this is easy to control for a genetic algorithm search as it is very dependent on the amount of mutation built into the algorithm. Though while the control is relatively simple, the amount of exploration required is hard to determine.

In multi-agent learning, one way to proceed is by utilizing an ε-greedy exploration method. Most of the time, with probability $(1-\varepsilon)$, the algorithm exploits current best behaviour, but once in a while, it explores randomly a behaviour with a small probability ε. An alternative approach is to use a Boltzmann "SoftMax" approach. The exploitation of the current one and exploration of an alternative behaviour are based on a parameter τ to balance exploration and exploitation. The probability P_i of a chosen behaviour μ_i among available behaviours $(\mu_1, \mu_2, \mu_3, ... \mu_j)$ in a state S denoted as:

$$P_i = \frac{e^{f(S,\mu_i)/_\tau}}{\sum_j (e^{f(S,\mu_j)/_\tau})}.$$

Another method, proposed by Lehman and Stanley (2011), utilizes a novelty approach to maintain diversity [13]. This approach utilizes how far apart one solution within a search space is from the other possible solutions. The novelty value is assigned to a given solution by the sparseness of behaviours within that section of the search space. The sparseness value ρ is an average distance to the k-nearest neighbours μ at a point x thus:

$$p(x) = \frac{1}{k}\sum_{i=0}^{k} dist(x, \mu_i).$$

This approach also has the advantage of reducing the risk of generating a solution to a local rather than a global goal.

3.3. Non-stationary behaviour

This is a particular problem when using genetic algorithms to develop control systems for distributed systems such as swarms. Each entity is free to evolve but it changes the relationship with all the other agents which, in turn, alternates the appropriate function for all other agents. It is possible to obtain a functioning solution for homogenous swarms, but there exists no robust solution for heterogeneous swarms.

3.4. The curse of dimensionality

Multi-agent systems that utilize learning mechanisms such as reinforcement learning which map the state to find the best behaviour suffer from the "curse of dimensionality". The calculated discrete states of the environment resulting from reinforcement learning in multi-agent systems grow exponentially with the increase of the number of agents. As the estimation increases for possible discrete state or state-action pairs, the complexity of the computation process lies in choosing the best behaviour or policy for the current state, which, of course, leads to longer computing time.

4. The addition of an epigenetic layer

An epigenetics layer is developed as a tool to assist an agent in responding to an environmental stimulus by modifying its phenotypic expression. It requires generating some type of regulatory means for an agent that receives an input from the environment (external stimulus) to use to regulate genotypes as a form of expression regulation. In order to achieve this, an algorithm, based on research work in biology, Sousa and Costa (2011) [14] proposed a method known as Epigenetic Algorithm (EpiAL). According to the proposed model, the interaction between the agent-environment is depicted in Fig. 2.

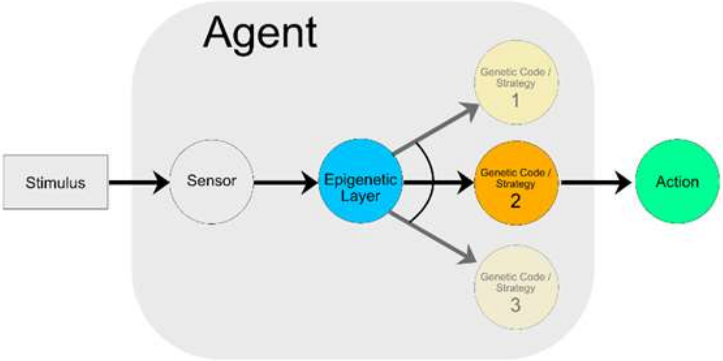
Agent
Stimulus
Sensor
Epigenetic Layer
Genetic Code / Strategy 1
Genetic Code / Strategy 2
Genetic Code / Strategy 3
Action

Figure 2 – EpiAL Conceptual Model (Sousa and Costa (2011))

An EpiAL model is composed of two fundamental entities – the agent and the environment. The Agent receives external stimulus from the environment. The stimulus is passed to an epigenetic layer which acts as a regulatory structure. The appropriate genetic codes are selected and regulated; the selected genetic codes are expressed, which modifies the current behaviour of the agent. After each cycle of the EpiAL algorithm, the performance of the behaviour is measured to calculate the relation between stimulus and the genetic codes defined as a methylation value. This methylation value is used to evaluate the weights of the epigenetic algorithm. This allowed representation of the regulatory function of epigenetic layer to be mapped into a mathematical model able to respond to a dynamic stimulus from the environment. Similar works in other studies also demonstrate the validity of this approach [15–17]. Fig. 3 represents the translation of the environment state to behavioural spaces [17]. The behavioural space contains the expressions composed of genes selected by the epigenetic layer.

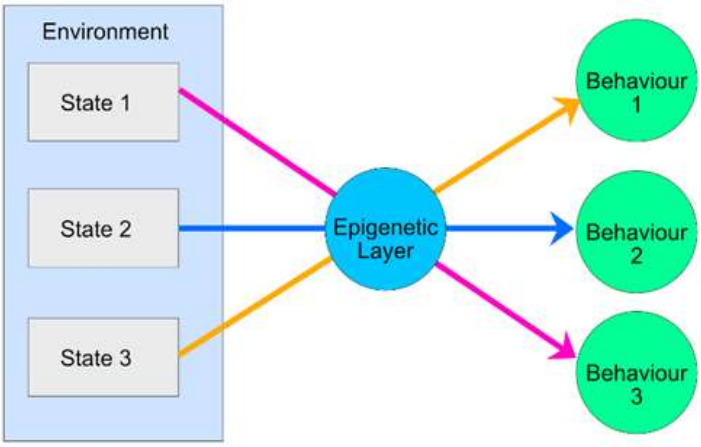

Figure 3 – Epigenetic layer map of the environment state to behavioural spaces translation

The regulatory function of epigenetic layer is based on the knowledge of sensed external stimulus from the environment. A temporal and spatial knowledge of the dynamic environment can be obtained through a trial-and-error approach. The behaviour in each interaction phase is evaluated and rewarded based on its performance in the current environmental state. The generated rewards are used as a basis of a relationship between individual agent's behaviours and environmental states. This generates a regulatory function for selecting a set of genetic codes that leads to a behaviour that creates a maximum reward for an environmental state. The model of this mechanism is depicted in Fig. 4.

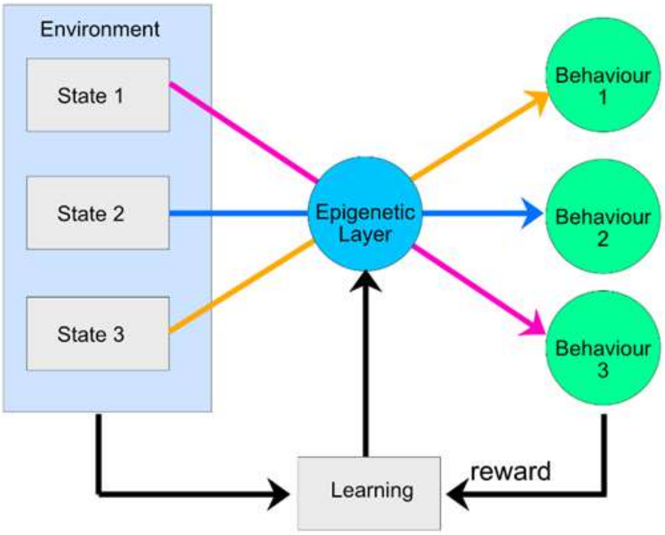

Figure 4 – Learning mechanism for epigenetic layer [18]

While this works quite well in practice, the main problem results in trying to evaluate the environment with a very limited sensor array. One way of improving this situation is to allow the system to learn as it progresses by embedding a state observer into the epigenetic layer. This leads to a process as displayed in Fig. 5.

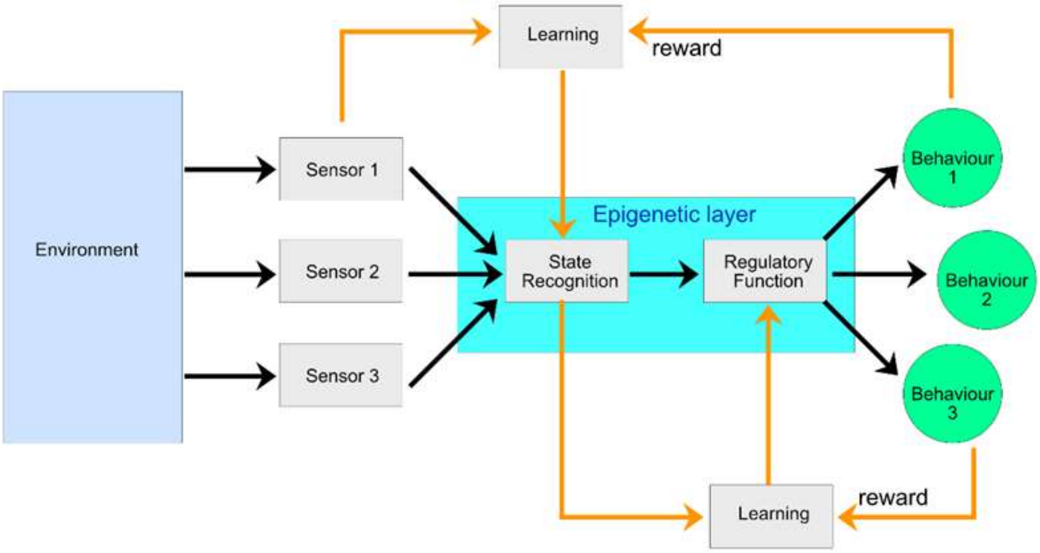

Figure 5 – Cascade learning for epigenetic layer [18]

5. Managing Alignment with an Evolving System

There are a number of possible approaches to this problem [19], but in practice, there are only two currently available methods. One of them is to imbed a human within the system to take action when some behaviour occurs that is not acceptable. This is akin to the role an airline pilot performs, where a largely autonomous system is constantly monitored, and if the skilled pilot is concerned, control is reverted to the human. The second approach is to have an external monitor who ensures that there will be no misalignment issues, which is akin to a regulator. This is a similar approach to that adopted by the launch crew of a space vehicle. The setup includes the expected value of each parameter being monitored, structural loads, flight dynamics, various temperatures, life indicators, etc. A track of the expected performance is generated before the flight and displayed to a specialist on a screen as a line with the indicated tolerance. As the flight progresses, there is generated a line of the actual parameter called a worm. The specialist watches the screen and if an abnormality occurs informs the mission controller. If a track exceeds the tolerance, the range safety officer may abort the mission. The problem with both these methods is that they are extremely slow in terms of computer speed, and they really work only on systems that, when deployed, retain the original behaviour.

There are two ways of evolving a system, either online or offline. In the offline case, a simulation is used to develop for the mission a workable solution, and once the solution is obtained, it is applied to the real-world problem, but this, of course, limits possible further improvements. This can, to some extent, be addressed by generating a digital twin that can be allowed to evolve until a significant advantage is achieved over the operational system when the newly developed variables and rules can be applied [20]. This evaluation is, of course, carried out by skilled human analysts and can cause delays in implementation. The alternative is to use online evolution. This has major advantages, particularly in a rapidly changing environment but to date, no one has achieved an acceptable working model with the guaranteed safety required, though there are numerous attempts to imbue artificial intelligence with an ethical response. The risks are still deemed too high to apply such a system in a working environment.

6. Alignment

Alignment is the term used to show that computer-based systems act in the best interests of humanity. In practice, this is quite hard to achieve, partly because it is very hard to quantify the best interests of humanity. A nation may honestly believe its weapon systems are in the interests of humanity as they prevent other nations or entities from taking action against them but other nations take a very different view. This is clearly represented by the US use of drones.

6.1. Potential in-built alignment for AI systems

The most common current approach to this problem is to try to build into systems acceptable ethics. Though Isaac Asimov’s three laws [21], namely: (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. All these can be seen as an interesting starting point, the problem is the definition of “harm” in the first law. It also offers no protection of sentient entities undefined as humans. In practice, there have been many attempts to define an ethical structure for artificial systems, one of the most significant being work done by the IEEE [22, 23]. The major problem with this approach is that there is no universal set of ethics that humans subscribe to. All communities, from families to nations, have an ethical structure that they believe to be superior. Thus it is extremely difficult, if not impossible, to create a set of ethical behaviours that all could accept. Even if it were possible, it would force the uniformity on human behaviour which would reduce valua ble diversity. This does not mean that it is impossible for artificial intelligence to act in the best human interest or at least in the best human interest of those who commission the system. Human society does this all the time, leading to individuals and nations being able to cooperate despite different ethical beliefs [24, 25].

6.2. How alignment might be achieved

One possible solution is to draw a distinction between the internal and external practices of an AI system. The main concerns with alignment only occur when the AI system interacts with the real world. As in human society, one's internal value system becomes relevant to other humans only at the interface. In human societies, we set a framework of laws that moderate this interaction. There are a number of such systems that perform quite successfully. The rules of engagement, deployed by military forces, are the ones that such system codified in the Geneva Conventions [26]. Another is the World Trade Organisation **[**27**]** which polices the rules of trade as established by the Marrakesh Declaration. It allows countries with vastly different ethical structures to conduct trade in a sustainable manner perceived to be fair to the participants. All such frameworks have elements that are regarded as compromises by different participants, but, what is most importantly, these systems work. Both these have in common is that they operate between entities with very different and often hostile world views. Perhaps the simplest example of how this might be applied is the self-driving car which, whatever its internal AI structure, is expected to adhere to the rules of the road. Of course, this does not define any actions outside the ruleset, which will depend on its internal ethics, but this is true for human drivers as well.

6.3. Producing a set of rules of engagement

It should be possible to set the appropriate rules of engagement for any interface using some form of directed deep learning. The learning algorithm would be directed to explore the outcome space, as in reinforcement learning. It would then be guided to evaluate which outcomes were not acceptable based on a number of criteria. It should then be possible after training for the program to determine which rules should apply for a particular unanticipated situation. This should lead to the possibility of judging which variables and values are acceptable within the AI system. From a swarm point of view, the boundary would depend on its nature. For a homogeneous swarm, the rule-driven boundary would be expected to be where the swarm interacted with the environment, while for a heterogeneous swarm, there would also be expected to be at the boundary between the agents.

7. Conclusion

This paper should not be perceived as a solution to the alignment problem. At best, it is an attempt to explore the problem and propose possible solutions. Much more refinement will be required before the solution proposed is proven to be valid. A great deal of further research will also be required. However, it does offer an alternative to the building of a global AI ethical standard which looks to be impossible.

REFERENCES

1. URL: https://www.bbc.com/news/technology-30290540.
2. URL: https://www.cser.ac.uk.
3. URL: https://futureoflife.org/.
4. Page J., Bain M., Mukhlish F. The risks of low level narrow artificial intelligence. *IEEE 2018 International Conference on Intelligence and Safety for Robotics (ISR).* 2018. (24–27 Aug. 2018). URL: https://ieeexplore.ieee.org/xpl/conhome/8517202/proceeding.

5. Nechayev S. Catechism of a Revolutionary. 1869. URL: https://www.marxists.org/subject/anarchism/nechayev/catechism.htm.
6. Leier M. Bakunin: The Creative Passion. Seven Stories Press, 2006. 384 p.
7. Reynolds C.W. Flocks, Herds and Schools: A Distributed Behavioral Model. *ACM SIGGRAPH Computer Graphics*. 1987. Issue 21 (4). P. 25–34.
8. Liu H., Gegov A., Cocea M. Rule Based Systems: A Granular Computing Perspective. *Granular Computing*. 2016. Vol. 1. P. 259–274. URL: https://link.springer.com/article/10.1007%2Fs41066-016-0021-6.
9. Goldberg D.E., Holland H.J. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, Mass: Addison-Wesley Pub. Co, 1988. Vol. 3. P. 95–99. URL: https://link.springer.com/article/10.1023/A:1022602019183.
10. Holland J.H. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. MIT Press, Cambridge MA, 1992. 232 p.
11. Nolfi S., Bongard J., Husbands P., Floreano D. Evolutionary Robotics. Springer Handbook of Robotics / Bruno Siciliano, Oussama Khatib (ed.). Cham: Springer International Publishing, 2016. P. 2035–2068.
12. Darwin C.R. The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Edited by John Murray. 6th edition. London: Eleventh thousand, 1872.
13. Joel L., Stanley K.O. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation*. 2011. Issue 19 (2). P. 189–223.
14. Sousa J.A.B., Costa E. Designing an Epigenetic Approach in Artificial Life: The EpiAL Model. *Agents and Artificial Intelligence.* 2011. Issue 129. P. 78–90. URL: http://link.springer.com/10.1007/978-3-642-19890-8_6.
15. La Cava W., Spector L. Inheritable Epigenetics in Genetic Programming. Genetic Programming Theory and Practice XII / Rick Riolo, William P. Worzel, Mark Kotanchek (ed.). Springer International Publishing, 2015. 3751 p.
16. Tanev I., Yuta K. Epigenetic Programming: Genetic Programming Incorporating Epigenetic Learning through Modification of Histones. *Information Sciences.* 2008. Issue 178 (23). P. 4469–4481.
17. Sathish P., Gray A., Kille P. The Epigenetic Algorithm. *Proc. of the IEEE Congress on Evolutionary Computation, CEC 2008* (Hong Kong, China, June, 2008). Hong Kong, 2008. P. 3228–3236.
18. Mukhlish F., Page J., Bain M. Evolutionary-learning framework: improving automatic swarm robotics design. *International Journal of Intelligent Unmanned Systems*. 2018. Vol. 6, N 4. P. 197–215.
19. Page J. Mentoring: A solution to the AI ethics dilemma? *International Conference on Artificial Intelligence. Information Processing and Cloud Computing* (Sanya, China, 19–21 December 2019). Sanya, China, 2019. URL: http://www.aiipcc.org/2019/KSCn.aspx.
20. Shaw K., Fruhlinger J. What is a digital twin and why it's important to IoT. Network World, 2019. URL: https://www.networkworld.com/article/3280225/what-is-digital-twin-technology-and-why-it-matters. html.
21. Asimov I. Runaround. New York: Street & Smith Publications, 1942.
22. IEEE Ethically Aligned Design An ongoing IEEE initiative Version 2. 2017.
23. IEEE ETHICALLY ALIGNED DESIGN First Edition. 2018. URL: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf.
24. Awad E., Dsouza S., Kim R., Schulz J., Henrich J., Shariff A., Bonnefon J.-F., Rahwan I. The Moral Machine experiment. *Nature*. 2018. Vol. 563, N 7729. P. 59–64.
25. OhEigeartaigh S.S., Whittlestone J., Liu Y., Zeng Y., Liu Z. Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governance. Philos Technol, Springer Link, 2020. DOI: 10.1007/s13347-020-00402-x.
26. Australian Law Geneva Conventions Act 1957. URL: https://ihl-databases.icrc.org/applic/ihl/ihl-nat.nsf/4fba3fefb860824b41256486004ad097/8a4aaee5c2dc9f88c1256b6d00303244?openDocument.
27. World Trade Organization Marrakesh Agreement Establishing the World Trade Organization from April 15 1994. URL: https://www.trade.gov/trade-guide-marrakesh-agreement-establishing-wto.

Стаття надійшла в редакцію 10.01.2022