The problem of AI identity

Soenke Ziesche
Independent researcher
Delhi
India
soenke.ziesche@gmail.com

Roman V. Yampolskiy
Speed School of Engineering
University of Louisville
USA
roman.yampolskiy@louisville.edu

Abstract

The problem of personal identity is a longstanding philosophical topic albeit without final consensus. In this article the somewhat similar problem of AI identity is discussed, which has not gained much traction yet, although this investigation is increasingly relevant for different fields, such as ownership issues, personhood of AI, AI welfare, brain–machine interfaces, the distinction between singletons and multi-agent systems as well as to potentially support finding a solution to the problem of personal identity. The AI identity problem analyses the criteria for two AIs to be considered the same at different points in time. Two approaches to tackle the problem are proposed: One is based on the personal identity problem and the concept of computational irreducibility, while the other one applies multi-factor authentication to the AI identity problem. Also, a range of scenarios is examined regarding AI identity, such as replication, fission, fusion, switch off, resurrection, change of hardware, transition from non-sentient to sentient, journey to the past, offspring and identity change.

Keywords: AI identity, personal identity, computational irreducibility, multi-factor authentication

Introduction

Identity, which can be also referred to as sameness, as a philosophical concept is about a relation between two or more things, which is true when the things are the same at different points in time. One of the challenges is to define, which properties of the concerned things are considered for sameness. For example, the sameness of the atomic composition of things is according to some philosophical positions not necessary for identity as it is discussed already in the ancient Ship of Theseus thought experiment (Plutarch, 75). The AI identity problem explores the sameness for two or more AIs at different points in time, and also here the basic question is: What properties of the concerned AIs are to be compared to establish sameness?

One identity problem, which has been examined at length, is the problem of personal identity, which addresses the question what it is that defines the continuity of a person over time (e.g., Olson, 2010, for an overview). Although no final consensus for this problem has been reached, we propose to commence the endeavour towards AI identity by looking at personal identity. Humans and AIs have in common that they reach certain states over time

by processing information while applying intelligence, and this ability appears to be a critical component for both to define their identity over time as will be outlined below.

Whereas humans also have the feature of being sentient, this is not clear for AIs, and actually by many not even considered at all. If (some) AIs were to be sentient, they would be, just like humans, a subset of the universe of minds (Yampolskiy, 2014). It has been argued that there is a non-zero probability for the existence of such sentient AIs, thus, their welfare deserves critical consideration, which includes the prevention of suffering and of deletion of such minds (Bostrom et al., 2018; Ziesche & Yampolskiy, 2018). While we look here at both non-sentient and potentially sentient AIs, the question of identity is especially relevant for sentient AIs as a prerequisite to address their individual welfare.

This article is structured as follows: First, we motivate the AI identity problem by outlining several fields of application. This is followed by the first proposed approach to tackle the problem, which includes an introduction to the personal identity problem. Thereafter the second proposed approach, based on multi-factor authentication, is outlined. We then examine a variety of scenarios. In the end, a summary and an outlook for future work is presented.

Motivation

Why is it critical to explore the AI identity problem? We have identified several areas of interest, which also shows that we aim to look at the problem from a wide lens, which includes philosophical aspects in particular:

1. It is relevant for legal matters to determine the criteria when an AI remains the same because only then ownerships (e.g., Yampolskiy, 2022; Peng et al., 2022) and patents (e.g., Fujii & Managi, 2018) for AIs may be claimed and tracked over time.[1]
2. A different dimension compared with ownership of AI would be personhood of AI, which is a topic of debates (e.g., Gunkel & Wales, 2021) and which, if granted, would also assign accountability to AIs, e.g., related to business relationships with them. This in turn would necessitate clarity about their identity over time.
3. Yet a further dimension would be if sentient AIs exist or come into existence. The identity of such AIs over time is relevant to study for various reasons, not only to ensure their welfare, as indicated above (Ziesche & Yampolskiy, 2018), but, e.g., also for future scenarios of relationships such as friendship or even marriage with sentient AIs.
4. Work is ongoing towards brain-machine interfaces involving AI (e.g., Zhang et al., 2020) and future scenarios may envisage uploaded human minds merged with AIs. Both cases also require clarity about the identity of the involved AI over time.
5. AIs may also collaborate as multi-agent systems, which may be for an observer undistinguishable from a singleton AI (e.g., Bostrom, 2006). Insights to the AI identity

[1] However, Yampolskiy (2022) describes that establishing AI ownership is confronted by a whole range of challenges since advanced AIs are unexplainable, unpredictable, uncontrollable, potentially capable of recursive self-modification as well as easy to steal and to obfuscate. For conceivable future scenarios that AIs may be granted legal personhood or other freedom rights or if sentience of certain AIs can be confirmed, ownership of those AIs would even likely be illegal.

problem would help to figure out whether it is a society of agents or a single agent we are dealing with.

6. Finally, this research may help solve problem of personal identity. Many debates about the problem of personal identity get complicated by components, such as consciousness, which are beyond the realm of contemporary science. In contrast, AI systems, i.e., computational processes, are more open to methods of contemporary science.

First approach

Given the possible similarities between the AI identity problem and the personal identity problem and the fact that the latter one has been researched for centuries, we introduce it here and examine what aspects can be transferred to the AI identity problem.

The personal identity problem

As it is outlined in an overview about the preservation of personal identity by Yampolskiy and Ziesche (2018) various views on personal identity have been developed, yet the current prevalent view is called psychological continuity, which can be traced back to Locke (1694). Of the more recent approaches the one by Shoemaker (1984) appears applicable to both, the personal as well as for the AI identity problem. He distinguishes between psychological connectedness and psychological continuity. A person is psychologically connected with a person in the past if she or he is now in psychological states *because of* psychological states she or he was in in the past. In other words, there is a causal relation between these psychological states. According to Shoemaker (1984) there is psychological continuity between persons at different points in time when the psychological states at a later point relate to those at an earlier point by a chain of psychological connections.

The topic of personal identity has gained recently new relevance since potential possibilities are discussed to transfer human minds to different substrates. For such, yet still theoretical scenarios some sub-scenarios are distinguished by Yampolskiy and Ziesche (2018), such as fission, fusion and resurrection of human minds. Fission is examined by Parfit (1984) in detail and is for humans rather a topic of thought experiments, yet for AIs due to their copyability simple to realise. In this regard another concept is relevant here, which is the so-called “multiple-occupancy view” (e.g., Noonan, 2003). According to this approach, the post-fission persons existed already prior to fission. This can be illustrated by a railroad track, which forks, potentially multiple times. While there is one track only, it figuratively overlaps all the branches to come.

Proposed definition

We suggest exploring whether Shoemaker’s (1984) notions of psychological connectedness and psychological continuity can be adapted towards AI identity.

This leads us to propose the following definition: If the state of a current AI has been caused by a state of one specific AI in the past, then there is a chain of connections between these

AIs, then there is continuity between these AIs, then there is identity between these AIs at different points in time.

It is important to clarify what is not required to ensure identity: Identity between two AIs does not mean that the two AIs remain identical over time since we have seen that this is not a requirement for personal identity either. To give two examples: 1) There can be still AI identity over time when the AI has (significantly) increased its knowledge and performance over time. 2) There can be still AI identity over time when the AI takes a "treacherous turn“ (Bostrom, 2014).

Concept of computational irreducibility

For the verification of this definition, we introduce a method, which harnesses the prevalent computer hardware substrate of AIs, i.e., the fact that computational processes are easier to formalize than neuronal ones. We propose to apply the concept of computational irreducibility (Wolfram, 2002) and argue that two AIs are identical if the latest one cannot be produced by any shortcut, but has to be computed from the original one. This can be further refined by enumerating the space of AIs as it has been done, for example, for the space of minds by Yampolskiy (2014). If we assign an integer to any AI, we can map it to states of specific cellular automata. Then, according to Wolfram (2002), some of those states of cellular automata are connected computationally, but there is only AI identity between two states if the latest can be attained from the former without performing intermediate state computations.

For symbolic AIs, which was the dominant paradigm until the 1990s, such causal relations between AIs at different points in time, thus, transitions between states of cellular automata can likely be shown as well as if the latest AI cannot be produced by any shortcut according to the concept of computational irreducibility. Yet, this appears challenging for sub-symbolic, i.e., contemporary, AIs in the light of unexplainability and incomprehensibility (Yampolskiy, 2020). However, there have been attempts in this regard (e.g., Mordvintsev et al., 2020).

Therefore, when it comes to the verification of the identity between two AIs at different points in time according to this approach, we may have to distinguish between two categories: The first category comprises all pairs of AIs at different times, whose identity is verifiable by humans, while in the second category are those pairs of AIs, whose identity is not verifiable by humans due to unexplainability and incomprehensibility. Nevertheless, the above definition of AI identity may still make sense, conceding that verifiability by humans may not be a relevant criterion.

Second approach

The second approach comes from a different angle, which is to harness established authentication methods for the verification of AI identity.

Multi-factor authentication - Introduction

The necessity for authentication to get access to electronic devices or services is ubiquitous nowadays. Since attempts towards unauthorized access are also widespread and getting more sophisticated, authentication systems have moved from single-factor to multi-factor authentication, of which six factors or categories are introduced here (e.g., Ometov et al., 2018):

- **Knowledge factor**: Something the to-be-identified knows, i.e., usually a password or a PIN.

- **Physical biometric factor**: Something the to-be-identified inherits, which is biometric, but static, e.g. fingerprint or iris recognition.

- **Behavioral biometric factor**: Something the to-be-identified inherits, which is biometric, but dynamic, e.g. gait analysis or mouse use characteristics (e.g., Yampolskiy & Govindaraju, 2008).

- **Ownership factor**: Something the to-be-identified owns, i.e., a physical item, such as a bank cards.

- **Location factor**: This factor is linked to the current location of the to-be-identified, which could be determined through a GPS signal or an IP address.

- **Guardian factor**: This factor does usually not appear in this list. It is based on a proposal by Buterin for social recovery wallets.[2] Similarly, for authentication a guardian could be involved, for example, friends, family members or institutions who testify for the to-be-identified.

Multi-factor authentication - Verification of AI identity

We are now exploring if such multi-factor authentication can also be applied towards verification of AI identity.

- **Knowledge factor**: Yampolskiy (2014) proposes a variant of a Turing Test to verify the identity of cloned minds, which could potentially be adjusted to approach the problem of AI identity. He describes a "interactive text-only communication" arrangement, which "proceeds by having the examiner (original mind) ask questions to the copy (cloned mind), questions which supposedly only the original mind would know answers to (…). Good questions would relate to personal preferences, secrets (passwords, etc.) as well as recent dreams." For the verification of AI identity, the arrangement of an altered Turing Test could be that an examiner asks questions to AIs at different points in time. The test would be passed if the examiner cannot distinguish between the two AIs by questioning them.

[2] https://vitalik.ca/general/2021/01/11/recovery.html

- **Physical biometric factor**: The only area we are aware of, for which the AI identity problem is currently being explored, is related to ownership verification where techniques are used resembling the physical biometric factor. For example, Peng et al. (2022) describe so-called model extraction attacks to steal successful machine and deep learning models by querying their application programming interfaces. They provide further an overview of the two categories of antidotes that have been developed to verify whether a model has been stolen (watermarking techniques as well as fingerprinting techniques) and introduce a novel fingerprinting approach.

- **Behavioral biometric factor:** Again the computer hardware substrate of AIs can be harnessed as usually some if not all of the previous "behaviour" of an AI has been recorded. Therefore the examined AI could be requested for authenticication to repeat certain unique previous behaviour. For example, a Dall-e could be asked to produce an image with a certain description and a different image generator may produce very different images, in terms of style.

- **Ownership factor**: This factor would be applicable for AIs with legal personhood (e.g., Gunkel & Wales, 2021). Those AIs would be able to own something, including unique items, which nobody else owns and which could be used for authentication.

- **Location factor**: The code of an AI is processed at a specific location, which includes decentralized AI. At any given time for any AI, which is connected to a power source, a location, usually within computer hardware, can be identified with a certain granularity where there is currently no other AI located. It could be as simple as IP address as described above for humans.

- **Guardian factor:** Also similarly as introduced above for humans, relevant agents, such as creators, users, trainers or other AIs, may or may not confirm that it is the same AI.

According to this method, the identity of an AI with an AIs at an earlier point in time is verified if all six factors of the multi-factor authentication have been passed.

Further scenarios

Based on the approaches above, we can establish that an AI, which just evolves over time, retains its identity, while the AI identity problem becomes partly trickier for the following scenarios, which are discussed below: Replication, fission, fusion, switch off, resurrection, change of hardware, transition from non-sentient to sentient, journey to the past, offspring and identity change. These scenarios illustrate that the AI identity problem has further facets than the personal identity problem since these scenarios are currently mostly impossible for humans and, if at all, only discussed regarding the personal identity problem in the light of emerging technologies.

Replication / Fission

Owing to their usual hardware substrate AIs can be easily replicated unlike humans, which increases the relevance of the discussion what this means for the identity of the involved AIs, while this is for humans for now of theoretical nature (Parfit, 1984). As the replicas will be independent, one replica will not cause the state of another replica. Likely, the replicas will quickly develop differently depending on their individual context, thus, their divergent input. Therefore, the replicas do not have continuity among each other, but each of them has continuity with the original. Therefore, there is AI identity between every replica AI and the original AI, but not among the AI replicas.

Fusion / Swarm

Fusion refers to a scenario where two or more AIs are merged into one. Like the scenario above the merged AI would have continuity with each of the originals and again the multiple-occupancy view can be applied, which has been introduced for the problem of personal identity above. A special, yet so far hypothetical case would be a merger between an AI and a human mind because of a brain–machine interface, as indicated under the motivations above. As also mentioned under the motivations, it could constitute a challenge to determine whether a certain AI is actually a merger between several AIs or a singleton AI.

When it comes to more than one identity according to the multiple-occupancy view, humans usually only think of diseases such as dissociative identity disorder and may not be able to conceive merged identities neither between AIs nor between humans and AIs. Nevertheless, if humans cannot conceive it does not mean that it is impossible.

What can be seen as another special case of fusion is swarm intelligence, composed of decentralized AI systems, yet acting collaboratively (e.g., Zhang et al., 2013). The inspiration often comes from biological systems. Examples are ant colonies, bee swarms, fish schools and birds flocks, which achieve a common goal more effectively than attempting it individually. While AIs are in a swarm this can be considered as fusion, potentially followed by fission as AI swarms may be a temporary arrangement only. Concerning AI identity, the same ideas apply as outlined above for fusion and fission.

Switch off

Switch off refers to scenarios when the AI has been removed from energy supply, but the code and the memory still exist.[3] The code and the memory could be preserved in different formats, including as hard copy, since above we declared the substrate as not relevant to AI identity.

[3] Concerning superintelligence there are discussions that it would prevent being switched off due to its assumed instrumental goal of self-preservation (Bostrom, 2014). However, this is not relevant here since we are looking at all types of AIs and even superintelligence may accept being switched off under certain circumstances, e.g., energy shortages or to “hibernate” during irrelevant or boring periods, given the prospect of to be switched on again with the same identity.

While being switched off such an AI should have, by definition, the identity of the AI before it was switched off, as opposed to losing its identity, because, as described below, it can be resurrected. If it was deprived off its identity while being switched off this would create problems regarding ownership, patents and also AI welfare in case of a sentient AI, which was switched off against its wish (Ziesche & Yampolskiy, 2018).

Resurrection / switch on again

While resurrection of humans is a longstanding yet impossible wish, resurrection of AIs is fairly straightforward and means to turn an AI on again after it had been switched off as described above. The duration for how long the AI had been switched off does not matter. While the fission scenario above can be compared with the "copy and paste" operation, resurrection resembles the "cut and paste" operation.

If there were no manipulations in between, the initial state of the AI after switching it on again is the same as the state the AI was in before it was switched off. Therefore, there is AI identity between a resurrected AI after a switch off and the AI before the switch off.

Change of hardware / substrate

During the existence of an AI its hardware could be significantly changed in various aspects. While from an evolution of technology point of view this would mostly concern upgrades, such as faster processors or more sophisticated sensors and actuators, we can in theory also consider downgrades. Potentially it is required for such a change to switch the AI system off and thereafter on again, as just described. Moreover, although for now AI is mostly implemented on the same type of computer hardware, a transfer of the AI to other substrates is also conceivable, in which the same type of computational operations continues to take place.

Nevertheless, AI identity should not be affected by any of such operations as the definition still applies that the first state of the AI after the change of hardware or substrate has been caused by the last state of the AI before the change of hardware or substrate. This is not only in accordance with the unchanged personal identity of humans who have, e.g., received glasses or a heart pacemaker or who lost a limb in an accident, but also with overall identity deliberations, e.g., about the Ship of Theseus, that the sameness of the underlying matter does not matter.

Transition from non-sentient to sentient (and vice versa)

As mentioned before, the specification of AI identity is especially relevant for sentient AIs. Yet, if sentient AIs were possible, also the sub-scenario is possible if not plausible that such AIs are not sentient from the moment they are created, but develop sentience over time, just as human babies do at early age. Since we assume, just as for living beings, that, if at all, sentience of AIs evolves without external manipulations the identity of the AI should not be affected during the transition from being non-sentient to being sentient.

In the future it may be possible to distinguish between AIs, which have the capacity to become sentient and those who do not have such capacity. To the former group Chalmers (2022) assigns a minimal moral status even before they are sentient, thus it would be helpful to categorize such AIs by means of their identity.

Furthermore, the reverse scenario is conceivable that a sentient AI transforms (again) into a non-sentient one, as it happens for humans in certain coma conditions. For example, if an AI endures unbearable suffering and is at the same time capable to turn its sentience off, it may as well do so. Also, for this transition the identity of the AI should not be affected.

Journey to the past

Another occasional human desire, yet unfeasible to implement is to go back in time (and to potentially revise certain actions). Also, this is an undertaking, which can be realized for AIs rather straightforwardly. The AI system just needs to be reset to a state, in which it has been in the past and of which records are likely available, which will guide the reset procedure. Such undertaking may be motivated in practice to evaluate how an AI would perform with different inputs and in different contexts.

Similarly, to the replicas in the fission scenario, the AI will likely develop differently and may not reach that state again, from which the journey to the past was initiated. Nevertheless, such an AI keeps its identity as it has gone back to an exact state, in which it was before. The subsequent branching off does not affect the identity since also in the new branch one state of the AI causes the following one.

Offspring

Scenarios are conceivable that AIs, through autogamy, create other AIs, which have to some extent been implemented already (e.g., Zoph et al., 2018). Such “children AIs” do not have the same identity as their “parents” since the offspring contains, according to both evolutionary theory and evolutionary computation, also random bits, which are not in the parent, thus, not computed by it.

Identity change

Since we discussed several scenarios where the AI identity is kept, we should also look at cases when the identity of an AI changes: The identity of an AI can be changed if there are external manipulations to the AI, which converts the AI to a state, which is not exclusively caused by a previous state of the AI. The external manipulation could be carried out by humans or other AIs, desired or undesired, e.g., through hacking.

It must be noted that this is not the case for all external manipulations of the AI because in the course of learning and updating AIs are, similar to humans, frequently exposed to external influences. Yet, only severe manipulations cause an interruption of continuity of AIs over time, thus, an identity change. In this case, the concept of computational irreducibility does not apply between the previous and the new state of the AI.

Lastly, we exclude, the possibility that an AI changes its own identity. Even if an AI is capable of modifying its own source code (significantly), this would still mean, according to our definition, that the new state of the AI has been caused by a state of that AI in the past.

Summary and future work

We have motivated the relevance of the AI identity problem for a variety of fields, ranging from legal issues, personhood of AI, AI welfare, brain–machine interfaces, the distinction between singletons and multi-agent systems, to supporting a solution to the problem of personal identity. Nevertheless, the issue of AI identity has hardly been examined yet in a comprehensive manner.

We suggested two approaches towards the AI identity problem: First, we summarised the status of the problem of personal identity and proposed an adjusted definition for AI identity, based on causal relations, thus, connectiveness, thus, continuity between AIs at different points in time. As a method of verification, we proposed to map AIs to states of cellular automata and to apply the concept of computational irreducibility to the transition from one state to another. Secondly, we suggested using multi-factor authentication for the verification of AI identity at different points in time based on the six factors knowledge factor, two biometric factors, ownership factor, location factor and guardian factor.

We tested this definition for scenarios such as replication, fission, fusion, switch off, resurrection, change of hardware, transition from non-sentient to sentient, journey to the past as well as offspring. And we also discussed how the identity of an AI may change.

Overall, this article aims to provide initial propositions to what appears to be a complex, yet relevant field of research. Therefore, due to the significance of the AI identity problem further work on the outlined definition and verification approaches is recommended.

In addition, we suggest considering applying the multi-factor authentication approach to the philosophical personal identity problem, which, according to our knowledge, has not been attempted yet.

Moreover, the focus in this article was on the Western philosophy of mind. However, not only for the purpose of inclusivity it appears promising for the AI identity problem to look in the future also at approaches how the mind is seen in Eastern philosophy. While the main distinction between dualism and monism has emerged in both Western and Eastern philosophy, also another doctrine has been developed in Eastern philosophy, which has no counterpart in Western philosophy: In the Buddhist philosophy of mind the term anatta stands for "non-self" and "holds that the notion of an unchanging permanent self is a fiction and has no reality" (Morris, 2006, p.51). Instead, a (sentient) being is defined by five so-called skandhas, which are form, sensations, perceptions, mental activity or formations and consciousness. It is beyond the scope of this article to explore whether this approach enables a more precise or even a different solution to the problem of AI identity and is, thus, also recommended for future work.

References

Bostrom, N. (2006). What is a Singleton? *Linguistic and Philosophical Investigations, Vol. 5, No. 2*, 48-54.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK.

Bostrom, N., Dafoe, A., & Flynn, C. (2018). Public policy and superintelligent AI: a vector field approach. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*.

Chalmers, D. (2022). Reality+: Virtual Worlds and the Problems of Philosophy. Allen Lane: London.

Fujii, H., & Managi, S. (2018). Trends and priority shifts in artificial intelligence technology invention: A global patent analysis. *Economic Analysis and Policy*, *58*, 60-69.

Gunkel, D. J., & Wales, J. J. (2021). Debate: What is Personhood in the Age of AI?. *AI & SOCIETY*, *36*(2), 473-486.

Locke, J. (1694). Essay Concerning Human Understanding.

Mordvintsev, A., Randazzo, E., Niklasson, E., & Levin, M. (2020). Growing neural cellular automata. *Distill*, *5*(2), e23.

Morris, B. (2006). *Religion and anthropology: A critical introduction*. Cambridge University Press.

Noonan, H. (2003). *Personal Identity*, second edition, London: Routledge.

Olson, E. T. (2010). *Personal identity*. Stanford Encyclopaedia of Philosophy.

Ometov, A., Bezzateev, S., Mäkitalo, N., Andreev, S., Mikkonen, T., & Koucheryavy, Y. (2018). Multi-factor authentication: A survey. *Cryptography*, *2*(1), 1.

Parfit, D. (1984). *Reasons and persons*. OUP Oxford.

Peng, Z., Li, S., Chen, G., Zhang, C., Zhu, H., & Xue, M. (2022). Fingerprinting Deep Neural Networks Globally via Universal Adversarial Perturbations. *arXiv preprint arXiv:2202.08602*.

Plutarch (75). *Theseus*.

Shoemaker, S. (1984). Personal Identity: A Materialist's Account. In Shoemaker and Swinburne, *Personal Identity*, Oxford: Blackwell.

Wolfram, S. (2002). *A new kind of science*. Champaign: Wolfram media.

Yampolskiy, R. V. (2014). The universe of minds. *arXiv preprint arXiv:1410.0369*.

Yampolskiy, R. V. (2020). Unexplainability and Incomprehensibility of AI. Journal of Artificial Intelligence and Consciousness, 7(02), 277-291.

Yampolskiy, R. V. (2022). Unownability of AI: Why Legal Ownership of Artificial Intelligence is Hard.

Yampolskiy, R. V., & Govindaraju, V. (2008). Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, *1*(1), 81-113.

Yampolskiy, R. V. & Ziesche, S. (2018). Preservation of personal identity—A survey of technological and philosophical scenarios. In *Death and Anti-Death*, ed. C. Tandy, Volume 16: 345-374. Ann Arbor: Ria University Press.

Zhang, X., Ma, Z., Zheng, H., Li, T., Chen, K., Wang, X., ... & Lin, H. (2020). The combination of brain-computer interfaces and artificial intelligence: applications and challenges. *Annals of Translational Medicine*, *8*(11).

Zhang, Y., Agarwal, P., Bhatnagar, V., Balochian, S., & Yan, J. (2013). Swarm intelligence and its applications. *The Scientific World Journal*, *2013*.

Ziesche, S. & Yampolskiy, R. V. (2018). Towards AI Welfare Science and Policies. *Special Issue "Artificial Superintelligence: Coordination & Strategy" of Big Data and Cognitive Computing, 3(1):2*.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8697-8710). https://openaccess.thecvf.com/content_cvpr_2018/papers/Zoph_Learning_Transferable_Architectures_CVPR_2018_paper.pdf