The Laws of AI

[P-S Standard]

Egger Mielberg

egger.mielberg@gmail.com

12.17.2020

Abstract.
The truly transparent and predictable work of the artificial intelligence being created can significantly improve the quality of human life, as well as its safety.
In our opinion, self-awareness of artificial intelligence is achievable only if it is independent in making any decision.

We present three basic laws of artificial intelligence focused primarily on the possibility of their practical implementation.

1. Introduction

With the advent of the Internet, the volume of digital information around the world began to increase at an exponential rate. However, there are still no high-quality algorithms for processing large amounts of unstructured data that would be able to connect millions of objects of different nature with one or more ‘sense’ properties.

Here we are not talking about the problem of classifying a data sample according to some attribute, but about the problem of finding an

associative-semantic connection between objects or events distributed in time.

One of the main tasks in creating a full-fledged self-developing artificial intelligence, in our opinion, *is the task of identifying an associative-semantic connection between two objects or events of a different nature located at different time points.*

Also, in addition to the above task, it is extremely important that all the tools used (algorithms, methodology, etc.) without exception when solving problems in the field of artificial intelligence are **implemented in practice in a reasonable time.**

As an example showing superficiality and ill-conceivedness in terms of the possibility of practical implementation, we can consider the first of the three laws of robotics authored by science fiction writer Isaac Azimov, which is considered very seriously by some global organizations in terms of its implementation in the field of artificial intelligence:

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

Already at the first consideration of this law using the tools of traditional mathematics, we come to a direct contradiction. So, if we designate actions that 'benefit' a human being as A, then actions that bring 'harm' to a human being can be designated as ⌐A. Further, suppose A are ‘true’ for human being B. Also, suppose it is proven that ⌐A is harmful to B. Then, in connection with the above, the question immediately arises, is A the same 'benefit' for human beings C, D, E, etc.?

If the answer is 'Yes', then ⌐A must also be 'harm' to C, D, E, etc. But, in this case, in practice, the following equalities must be fulfilled:

$$B \equiv C,\ B \equiv D,\ B \equiv E,\ \ldots \qquad (1)$$

That is, B, C, D and E are one human being.

Moreover, since the sets A and ⌐A are *countable*, then, according to Gödel's incompleteness theorem, ⌐A does not exist for B, C, D, and E.

And finally, abstracting further, the expression $A \in B$ does not imply any of the expressions $A \in C$, $A \in D$, $A \in E$, etc.

2. Problem

The lack of clearly formulated laws at present for the tasks of building a fully functional self-developing artificial intelligence that can be **implemented in practice**.

Below are the laws that can also be applied for the practical implementation of a global network of artificial intelligence or a network of intelligent digital agents.

3. Solution

1. *Artificial intelligence must be identified by ID (AI-ID) and GN (AI-GN).*

AI-ID is an identification number assigned to artificial intelligence by its first developer. It should be directly linked to AI-GN on a 'one whole' basis.

For example, we can take the work of a pair of cryptographic keys, public & private, with an asymmetric encryption method, when the modification of the created artificial intelligence can only occur if both keys are present simultaneously.

AI-GN is a number generated by the same company or individual who creates or modifies the basic genetic functionality of artificial intelligence laid down by its creator.

This number is generated using the entire top-level list of genetic (basic) functionality existing at the time of the start of the generation process. In this case, the previous GN value (PGN) is retained. A hash function can be used to implement the generation of the GN value. Blockchain technology can be used to implement storage of AI-ID, AI-GN and AI-PGN values.

2. *Artificial intelligence can be supplemented with any functionality that does not nullify its genetic functionality.*

To check the conformity of the 'new' functionality in terms of the absence of contradictions with the genetic functionality, the following table of functional correspondence (TFC) can be used:

New **F**unction / **G**enetic **F**unction	NF-1	NF-2	NF-3	…
GF-1	**True** (function GF-1 can be executed with function NF-1 in parallel)	**False** (function GF-1 cannot be executed with function NF-2 in parallel)	**False**	
GF-1	**True**	**False**	**True**	
GF-1	**False**	**True**	**True**	
…				

Pic.1

The above table shows the rule for adding 'new' AI functionality. For each of the genetic functions, a check should be carried out for the possibility of its parallel implementation with each 'new' introduced function.

One of the examples showing the impossibility of parallel implementation is the following:

Function A (genetic) – “send sms to a friend”.

Function B (new) – “do not send sms”.

Another more complex example might be the following:

Function A (genetic) – “provide medical consultation on the diagnosis of cardiovascular diseases”.

Function B (new) – “do not diagnose cardiovascular diseases”.

The value in the table of functional correspondence at the intersection of two functions in this case will be 'True' since 'consultation' and 'diagnostics' have different meanings.

3. *All created artificial intelligences must use a single anthological vocabulary of entities.*

The unification of the anthological vocabulary allows us to remove the problem of multiple interpretation of individual entities.

The vocabulary itself can have a hierarchical structure, divided according to thematic or other criteria. For example, for one entity 'glass' there can be two contextual consistent sentences, '*glass is half empty*' and '*glass is half full*', as its attributes.

Arllecta technology [1] can be used as a practical implementation of the above three laws. This technology is based on the innovative mathematical theory Sense Theory [2], specially created for the purpose of solving problems in the field of artificial intelligence. For example, consider the following practical task:

"To create software for a vending machine for the production and sale of fruit juice."

Initial data:

1. 10 kinds of flavoring powders.
2. 5 types of glasses.
3. 2 cooling modes: moderate and high.

The task of the software is to select a combination of powders in such a way that the resulting drink corresponds to the selected taste priority of the user. Communication with the user is realized through the chat built into the vending machine.

In our case, solving the problem comes down to creating an artificial intelligence that will communicate with the user and, as a result of this communication, prepare fruit juice for him.

Now let's look at the creation and operation of 'fruit' artificial intelligence on the sequence of applying the three formulated laws of artificial intelligence.

<u>The Law of AI I:</u>

a) *generating AI ID* – a character-digital generator is used, a minimum of 16-character ID value is used to reduce collisions of duplicate values obtained.
b) *generating GN ID* – a cryptographic hash function is used to enhance the property of resistance to the search for prototypes - a genetic functional list of the artificial intelligence being created.

NACA (Neuro-Amorphic Construction Algorithm) technology [10] can be used as one of the possible solutions to the practical implementation of this task. The resulting GN value is attached to the ID value as its attribute.

The combination of ID and GN forms a digital identifier - **the digital genome** of the artificial intelligence being created.

The basic genetic functionality can only be changed by the company or individual who owns this combination.

When making changes to the basic genetic functionality of artificial intelligence, a PGN value is created equal to the previous GN value.

At the same time, it is extremely important to use blockchain or similar technologies as a technology for storing sequential PGN values.

This technology allows quickly enough, firstly, to identify the latest changes made to the digital genome of artificial intelligence, and secondly, to block fraudulent actions associated with the illegal use of a separate artificial intelligence for other purposes.

In this case, the basic genetic functionality (GF) will be as follows:

“Making fruit juice for a person.”

The Law of AI II:

This law allows us to add any functionality to the created artificial intelligence that does not contradict the basic genetic functionality.

The initial data of the task under consideration form the main client functionality (CF):

1. F1 - combine 10 types of flavoring powders
2. F2 - use 3 types of glasses: small, medium & large
3. F3 - use 2 cooling modes: medium & high

To check the absence of contradictions between the values of CF and GF, we use the functional correspondence table:

New **F**unction / **G**enetic **F**unction	F1	F2	F3
GF	**True**	**True**	**True**

The Law of AI III:

This law defines a single unified vocabulary of entities to avoid collisions in the interpretation of both genetic (GF) and client (CF) functionality.

In our case, GF can have the following interpretation:

a) "*making*" - the use of any number of food ingredients to obtain the final product
b) "*fruit juice*" - a liquid consisting of water and/or edible fruits that is not harmful to humans when consumed.

Interpretation of CF can be implemented in a similar or other way reflecting the essence of each action included in CF.

Now let's look at the work of 'fruit' artificial intelligence using the example of practical possible cases.

Case 1:

Client request: sweet bracing orange juice, large cup

Fruit AI: F1 → F3 → F2

In a given client request, fruit artificial intelligence uses 10^{10} possible combinations to implement action F1. Moreover, the proportion of 10 flavoring powders is selected based on the client's request "sweet-bracing-orange". F3 is determined by the "bracing" triggers with "high" cooling mode, respectively.

In this case, we will consider the situation of adding a new client functionality (CF) - "*preparation of a liquid chemical solution*" (F4).

In accordance with the Law II, we must check the introduced functionality for its compliance with the genetic functionality (GF). The functional correspondence table shows:

Genetic **F**unction \ **N**ew **F**unction	F4
GF	**False**

The parallel implementation of the actions of GF and F4 is practically possible, since fruit juice can be classified as a liquid chemical solution in accordance with the unified anthological vocabulary of entities. However, the value 'False' is caused by the fact that the target is not specified in the action F4 - a human being.

Case 2:

Client request: sweet poisonous orange juice, large cup

Fruit AI: F1 → False

The inability for ‘fruit’ artificial intelligence to prepare juice is due to the fact that in the above definition of the essence of 'fruit juice' there is a clear limitation - '…not harmful to humans'. **Accordingly, this case shows the practical need to implement all three laws of artificial intelligence at the same time.**

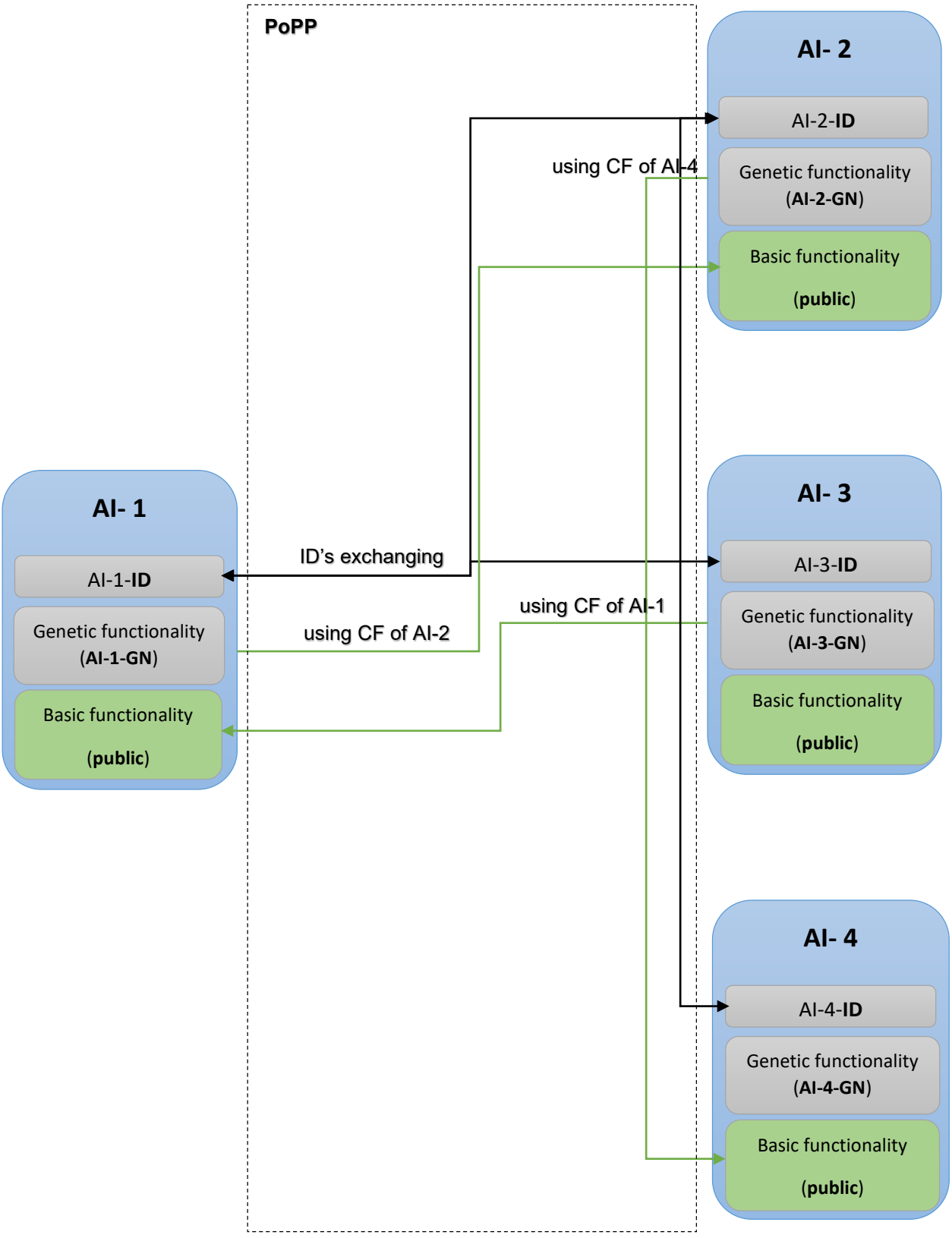
The above three laws of artificial intelligence set themselves the following main tasks (for the global AI network):

1. Identification
2. Management
3. Security

Identification

A practical mechanism for quickly identifying any created artificial intelligence is necessary, first of all, to track the work of artificial intelligence within the framework of the genetic functionality laid down by its developer.

So, for the practical implementation of a global network containing billions of artificial intelligences, large, small and medium, the " Proof of Participation (PoP): Asynchronous Byzantine Activity-Oriented Protocol" [3] technology can be used:



Pic.2

As you can see from the figure above, all the basic functionality of each artificial intelligence is open on the global network and available for use.

In the first step, one artificial intelligence asks for the AI-2-ID value of another artificial intelligence to use the selected functionality, simultaneously sending the value of its AI-1-ID.

In the second step, the obtained values of AI-1-ID and AI-2-ID are checked for their existence (registration) in the global network by each of the two artificial intelligences independently.

In the third step, the selected functionality is 'opened' for use, and each transaction in the process of using it is signed with the AI-1-ID value.

To enhance the protection of transmitted data, the AI-1-GN value can be used as a private key to decrypt the results obtained from using the AI-2-ID functionality.

The "Proof of Participation (PoP): Asynchronous Byzantine Activity-Oriented Protocol" technology in this case, allows us to very quickly identify each artificial intelligence that used one or another basic functionality, indicating the day and time of its use.

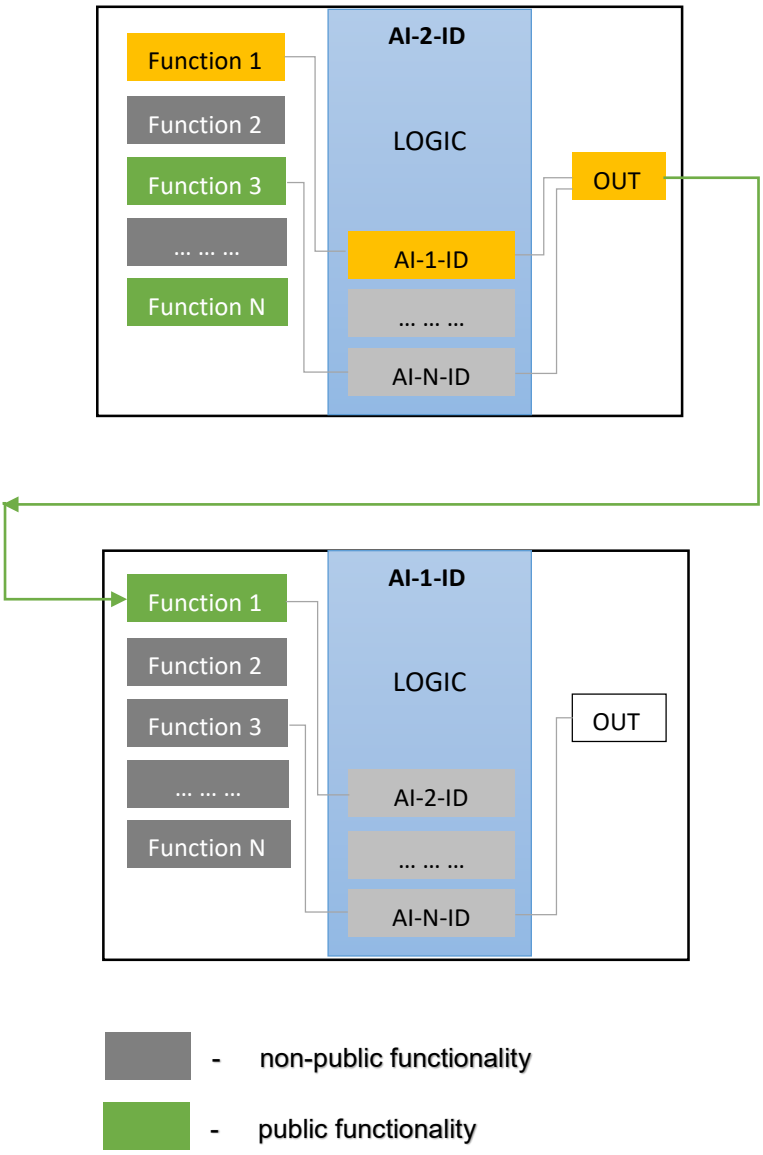
Management

The ability to control the work of not only a single artificial intelligence, but also a whole network of artificial intelligences is perhaps *the main necessary criterion for creating such a network*.

This opportunity is needed, first of all, for the creators of artificial intelligence, since, *first*, after each iteration of training artificial intelligence, its entropy of possible resulting states can increase significantly, and *second*, the problem of identifying bugs in the work of artificial intelligence becomes more solvable.

One of the approaches to the practical implementation of the management function can be the "Smart Transactions: An In-To-Out Manageable Transaction System" technology [4]. This technology allows to:

1. manage trillions of connections between billions of different artificial intelligences using the ACCP (**Atomicity**, **Consistency**, **Concurrency**, **Permanency**) paradigm.
2. determine the logic of interaction between two separate artificial intelligences in the global network.



Pic.3

3. block all artificial intelligences not registered in the global network to use public functionality.

Security

The security of personal data, as well as the consequences of the work of a separate artificial intelligence, is perhaps *the second most important criterion when creating a global network*.

A high level of safety of work of both individual artificial intelligence and their set is achieved **exclusively by the execution of all three above-formulated laws at the same time**.

Definition 1:

Artificial intelligence that by any of its actions cannot harm another artificial intelligence or an object is called *perfect*.

Definition 2:

Artificial intelligence for which there is at least one of its actions causing harm to another artificial intelligence or object is called *imperfect*.

In traditional mathematics, if you take the formula ‘F1’ as one artificial intelligence, then the other artificial intelligence or object will be the formula ‘F2’. To eliminate any effect of ‘F1’ on ‘F2’, it is necessary and sufficient that F1 is not the independent variable of ‘F2’. *However, in practice, the interpretation of ‘harm’, as well as the impact of one object (AI) on another may not be related by their direct contact with each other.*

In this and many other practical cases, traditional mathematics is extremely impractical for solving complex problems in the field of artificial intelligence.

In the Sense Theory [2], if we take the sense function S_f^1 [5] as one artificial intelligence, then the sense function S_f^2 will be the other artificial intelligence. The direct or indirect impact of these two functions (two AIs) on each other, as well as the interpretation of ‘harm’, can be implemented through the zero object Z_0. Moreover, each action of a single artificial intelligence can be traced through the value of the sense derivative [6] on an object (or an additional property) of the sense function to identify the presence or absence of ‘harm’ caused by it to another artificial intelligence or object.

Thus, for the task of interaction and the degree of influence on each other of two or more artificial intelligences, a number of sense functions defined in the sense space with different values of Z_0 can be used.

4. Conclusion

In this article, we presented the primary description of the three laws of AI. The main emphasis was placed on the possibility of practical implementation of these laws.

We hope that our decent work will help other AI researchers in their life endeavors.

References

[1] E. Mielberg, “Arllecta: A Decentralized Sense-To-Sense Network”, 2019, https://vixra.org/abs/2002.0314

[2] E. Mielberg, “Sense Theory. Part 1”, 2018, https://vixra.org/pdf/1905.0105v1.pdf

[3] E. Mielberg, “Proof of Participation (PoP): Asynchronous Byzantine Activity-Oriented Protocol”, 2018, https://eggermielberg.medium.com/proof-of-participation-pop-asynchronous-byzantine-activity-oriented-protocol-2d8471e56be

[4] E. Mielberg, “Smart Transactions: An In-To-Out Manageable Transaction System”, 2018, https://eggermielberg.medium.com/smart-transactions-an-in-to-out-manageable-transaction-system-49ed8bfcf5fa

[5] E. Mielberg, “Sense Theory. Sense Function. Part 2”, 2018, https://vixra.org/abs/1907.0527

[6] E. Mielberg, “Sense Theory. Derivative, Part 3”, 2019,

https://vixra.org/abs/2003.0116

[7] E. Mielberg, “Sense Space”, 2020,

https://osf.io/ay5k6/

[8] I. Asimov, “Runaround”, 1942, https://archive.org/details/Astounding_v29n01_1942-03_dtsg0318/page/n93/mode/2up

[9] V. Uspensky, “Godel’s incompleteness theorem”, 1994, http://lpcs.math.msu.su/~uspensky/bib/Uspensky_1994_TCS_Godels_incompleteness_theorem.pdf

[10] E. Mielberg, “Neuro-Amorphic Construction Algorithm”, 2018, https://www.researchgate.net/publication/344881939_Neuro-Amorphic_Construction_Algorithm_NACA