Stability of artificial intelligence controlled systems

Vaclav Jirovsky, Vaclav Jirovsky jn.

Faculty of Transportation Sciences

Czech Technical University in Prague
110 00 Prague, Czech Republic

ABSTRACT

The article deals with the problem of stability of system controlled by artificial intelligence. The model based on theory of structures is designed and later discussed.

Keywords: System of systems, artificial intelligence, structures of systems

INTRODUCTION

In the middle of 18th century the industrial revolution, major turning point in the history, caused transition of hand production methods to industrial production. Although mechanized factory relieves humans from heavy manual work transferring it to machines, the workers destroyed new machines in the belief that the machines were taking over their jobs. Today we are observing similar but different process – IT revolution, accompanied with deployment of the information technology in all areas of human doings. The transfer of "intellectual work" from people to machines is broadly welcomed, under belief that machines will do it better, with no errors, and faster. No one is protesting, no one is destroying computers, everyone gladly greets transfer of human privileges of thinking and decision making to machines.

New Artificial Intelligence[1] – the capability of machines to reason, communicate and make decisions is at the center of IT revolution (Jirovsky & Jirovsky jn., 2020). The development is moving ahead uncontrolled and even there are some lonely screams about danger, the marketing drown it out (Osborne, 2017), (Clifford, 2018). Some authors point out danger of accidents caused by AI implemented as part of control systems very loudly, some just points out what should be consequences of implementation of faulty program[2] (Perrow, 1999). Deploying AI is an ongoing process that holds tremendous promise as well as equally remarkable danger (Anon., 2018).

Consider growing implementation of systems using elements of AI and communication, each with other, we have to ask, what is the actual danger. The answer is simple – the system of systems becomes danger if the complex system becomes unstable.

MEANING OF THE STABILITY

The stability of a control is defined by many authors as the ability of system to provide a constrained output when a specific input is applied to the system. So we should expect, that if the system reaches the steady-state it remains in that state for that particular input. This is important property of a control system. We should say that such a system is **absolutely stable** and as could be found, the transfer function of the system will be strictly uniform. Such a system will stay in the defined state even the parameters of the system are changed.

Another situation will be in the case when the output variable and input variable relation are dependent on specific condition of the system that is defined by the parameter of the system. In this case the system could be called **conditionally stable system** where output is restricted by system parameters and could lead asymptotically to stable state. The most critical condition arises when output of the system oscillates in the response to change in the input variable or change in system parameters. Such a system we can call **marginally or critically stable** system.

All these behaviors can be found in the well-known basic structure of common controlled system, which is depicted on the Fig. 1. The state of environment is detected by sensors and system reacts by actuators to adapt, control, its preprogrammed behavior to state of the environment. In contrary, environment under effect of the actuators changes its state. But the new state could be again different from the state expected under influence of actuators. So, whole system will oscillate around the final state trying to find an equilibrium, where all components – system and environment will be stable. As had been mentioned, this situation is well known and described in countless number of articles.

[1] We should carefully point out that term „Artificial Intelligence“ is nothing more than marketing issue. Real intelligence is using intuition, emotion, morality or ethics. Replacing human ability of abstraction by machine learning is not enough.

[2] Actually we cannot say the AI program is faulty, because one of the basic features of the AI program is that the programmer (author) is not able to predict or estimate the next steps of the program – its behavior.

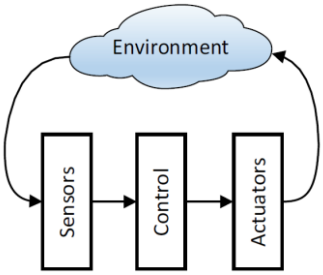

Fig. 1 Classical controlled system

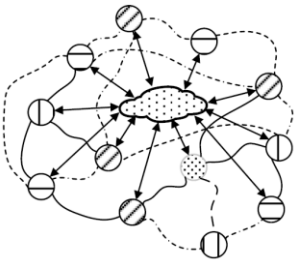

Fig. 2 Structure of interconnected systems and environment

What happens when elements of the artificial intelligence will be implanted into such simple control mechanism. The artificial intelligence element "in statu nascendi" will know just elementary mechanisms of reaction embedded during manufacturing process. The artificial intelligence will try to use these mechanisms and senses the reaction of the environment. There are several possibilities, which could happen, particularly similar to those related to the system without AI:

- there will be no response from the environment,
- there will be appropriate and expected reaction of the system and whole complex will reach stable state until next change in the environment or system,
- the response of the environment could be so drastic, that system will be destroyed,
- the action of the system will lead to the destruction of the environment.

In this case, from the ontological point of view, it is a substantive approach that does not look for systemic connections, but accepts the system as such, without studying its internal organization (Šmajs, 2008). A more complex situation arises in the case of a system of systems[3], where one of the decisive facts will be the structure of the interconnections among systems[4] and their internal mechanisms, which are represented by their external state projected into the overall state of the system of systems reacting to the environment. At the same time, the environment is influenced by the action of other connected systems – see Fig. 2, creating an environment for other systems. It finally leads to a very complex behavior of the whole system of systems[5].

From this point of view, we can describe such a structure of the interconnected elements depicted on Fig. 2 by mathematical formula (Jirovsky, 1974) as

$$\mathcal{H} \equiv \langle \mathcal{Z}^n, \mathcal{S}^p, L_m^{\vartheta} \rangle$$

where

[3] In this case we expect that systems creating system of systems are more less elementary automatons than another complex system.

[4] Moreover, this interconnection is subject to dynamic changes.

[5] Usually the stability of large systems is studied using Lyapunov functions and other mathematical means but we believe that such analysis will not be usable in the systems controlled by artificial intelligence or even self-learning program.

$\mathcal{H}$ is structure of system of systems formed by ordered triplet,
$\mathcal{Z}^n$ is n-dimensional space where system of systems exists,
$\mathcal{S}^p$ is a set of all possible states of systems in $\mathcal{Z}^n$, where $\mathcal{S}^p \in \mathcal{S}$ and $card(\mathcal{S}^p) = p$,
L_m^{ϑ} is local behavior function of the system in $\mathcal{Z}^n$, where m is a number of neighbors and ϑ is a neighbor index which is represented by ordered set of $\vartheta_1, \vartheta_2, \ldots \vartheta_n$ and which express relations to the other systems linked to this system.

The meaning of this formula could be easily demonstrated in 3-dimensional space, where all elements, in this case simple final state automatons, will be placed on Euclidean plane and their state will be projected to *z*-dimension. The result will be set of dots of different height describing current state of the structure of system of systems – see Fig. 3. The stability of the structure of automatons then could be defined e.g. as limits of the state diagram. In our example, it will be represented by planes on the bottom and on the top of 3-D graph.

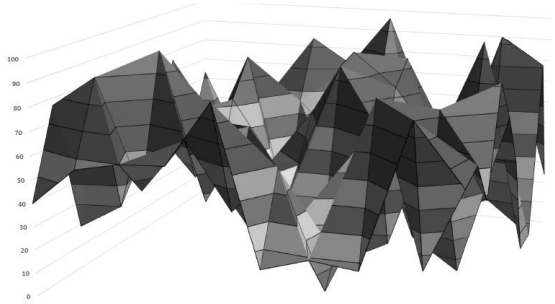

Fig. 3 3-D representation of state of structure of the automatons shown as collapsing surface

This simple situation could be extended into larger scale of system of systems and implementation of stochastic behavior in every element of the system of systems. Then we can define new structure called stochastically non-homogeneous structure $\upsilon \equiv \{\mathcal{Z}^m, \mathcal{S}^p, \mathcal{S}^0, P_v\}$, where behavior function of the elements will be given by stochastic matrix $P\upsilon = \{p_{ij}^v\}$ for any time t > 0 as

$$p_{ij}^v = Prob\left[L_v^t: \left(\mathcal{K}_{\mathcal{R}_A}\langle A, t\rangle \rightarrow s(A, t+1) = s_j\right)\right] \quad \text{and} \quad p_{00}^v = 1$$

where

$\mathcal{K}_{\mathcal{R}_A}\langle A, t\rangle$ is known configuration of the structure at time *t* if subset of neighbors is $\mathcal{R}_A$,
s_j is a new configuration in time t+1,
L_v^t is operator of configuration mapping, which could be understood as global meaning of L_m^{ϑ} ,

For excitation of the whole system of systems, it would be necessary to have probability p_{00}^v in time 0 equal to 1.

Furthermore, it should be shown that any structure $\{\mathcal{Z}^m, \mathcal{S}^p, \mathcal{S}^0, P_v\}$, where $\mathcal{S}^p$ is a set of all possible states of systems in $\mathcal{Z}^n$ and $\mathcal{S}^0$ is state of the system of the systems in $\mathcal{Z}^n$ in time t=0, will be equivalent to some of the Markov chains

(Aladyev, 1980). Thus, with respect to above we can say that behavior of stochastically non-homogeneous structure will be stable if equivalent Markov chain will be convergent.

Nevertheless, we only know $\mathcal{S}^0$, and the number of states at t = 0. However, the number of states in the time t >> 0 is unknown. Moreover, if the Markov chain will be the kind of Markov chain with absorption, some of the states in past will disappear from the set of states and after some time they could appear again. It means that $card(\mathcal{S}^p)$ will not be constant but it is dynamically changed over the course of time. It makes the solution of problem of stability in longer time difficult if not unsolvable in reasonable time. More details on general convergence of Markov chain could be found in (Suhov & Kelbert, 2008).

CONCLUSION

In 2006, Boardman and Sauser (Boardman & Sauser, 2006) state that the stability of a control system is defined as the ability of any system to provide a bounded output when a bounded input is applied to it. With regard to previous analysis, we can more specifically say that system is stable if Markov chain model of the system of systems will be convergent.

In the case when artificial intelligence will be implemented into control structure of the system in system of systems, the natural tendency of the system will be to achieve the optimal condition for its "survival". However, one of the important characteristic of the program implementing artificial intelligence is that the programmer itself is not able to predict next behavioral steps of the program. From this point of view, the system of systems behaves stochastically and as previous analysis had shown, the stability could be achieved only and only if the respective Markov chain model of the system of systems will be convergent.

In a such case we should switch from the mathematical proofs to psychological theories describing behavior of human groups, simply because the proofs are nearly impossible. Such a problem of "community behavior" had been studied by psychologist Irving Janis (Janis, 1971) who coined the term "groupthink", which refers to a set of dysfunctional decision processes used by highly cohesive groups, leading them to ignore alternative courses of action and to discourage irrationally the expression of disconfirming opinions (Jirovsky & Jirovsky jn., 2021). The more precise description of this process based on stochastic structures models is subject to further research.

REFERENCES

Aladyev, V. Z., 1980. *Mathematical Theory of Homogeneous Structures and their Applications.* Tallin: Valgus.

Anon., 2018. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation.* [Online]
Available at: https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf [Accessed January 2019].

Boardman, J. & Sauser, B., 2006. *System of Systems - the meaning of of.* s.l., s.n.

Clifford, C., 2018. *Elon Musk: 'Mark my words — A.I. is far more dangerous than nukes'.* [Online]
Available at: https://www.cnbc.com/2018/03/13/elon-musk-at-sxsw-a-i-is-more-

dangerous-than-nuclear-weapons.html
[Accessed October 2020].
Janis, I., 1971. Groupthink". *Psychology Today,* November.p. 43.
Jirovsky, V., 1974. *Teorie homogenních struktur a její aplikace v radioelektronice (Theory of homogenious structures and their appplication in the radioelectronics - in Czech),* Prague: Czech Technical University.
Jirovsky, V. & Jirovsky jn., V., 2020. Impact of disruptive technologies on the human attitude. *Advances in Intelligent Systems and Computing,* Volume 1018, pp. 84-89.
Jirovsky, V. & Jirovsky jn., V., 2021. *Can artificial intelligence be held responsible?*. s.l., Springer, pp. 605 - 610.
Osborne, H., 2017. *Stephen Hawking AI Warning: Artificial Intelligence Could Destroy Civilization.* [Online]
Available at: https://uk.news.yahoo.com/stephen-hawking-ai-warning-artificial-094355027.html
[Accessed August 2019].
Perrow, C., 1999. *Normal Accidents: Living with High-Risk Technologies.* Princeton: Princeton University Press.
Suhov, Y. & Kelbert, M., 2008. *Probability and Statistics by Example: Volume 2, Markov Chains: A Primer in Random Processes and Their Applications.* Cambridge: Cambridge University Press.
Šmajs, J., 2008. *Uvedení do evoluční ontologie (in Czech).* Brno: Masaryk University.