systema

2014 | Volume 2
Issue 2 | 61–70

ISSN 2305-6991
BCSSS

Performance-Based Control of Learning Agents and Self-fulfilling Reductionism

Yagmur Denizhan

Electrical and Electronics Engineering Dept., Bogazici University, Istanbul - Turkey, denizhan@boun.edu.tr, +90-212-3596850

Abstract: This paper presents a systemic analysis made in an attempt to explain why half a century after the prime years of cybernetics students started behaving as the reductionist cybernetic model of the mind would predict. It reveals that self-adaptation of human agents can constitute a longer-term feedback effect that vitiates the efficiency and operability of the performance-based control approach.

Keywords: Performance-based control; cybernetics; reductionism; adaptation

Cybernetics, right from its birth in mid 20th century has proposed a very characteristic way of modelling the functioning of living beings and particularly that of the human mind, giving birth to the new discipline of cognitive science (Dupuy, 2009). In the course of the years this reductionist conception that suggests a mechanistic model of the mind has been target to many criticisms and instigated a quest for alternative models.

In spite of the many decades and alternative views that intervened, I have come to observe in the first decade of the new century the emergence and proliferation of some behavioural patterns that are very much compatible with the reductionist model of the mind. Although I am being supplied many supporting examples from the business world and other domains of life, I will ground my arguments on first-hand observations I have been making during my academic career as a professor of electrical and electronics engineering.

1 Introduction

What led me to the line of thought underlying this article was a strange situation I encountered sometime in 2007 or 2008. It was a new attitude in my sophomore class that I never observed before during my (by then) 18 years' career. During the lectures whenever I asked some conceptual question in order to check the state of comprehension of the class, many students were returning rather incomprehensible bulks of concepts, not even in the form of a proper sentence; a behaviour one could expect from an inattentive school child who is all of a sudden asked to summarise what the teacher was talking about, but with the important difference that --as I could clearly see-- my students were listening to me and I was not even forcing them to answer. After observing several examples of such responses I deciphered the underlying algorithm. Instead of trying to understand the meaning of my question, searching for a proper answer within their newly acquired body of knowledge and then expressing the outcome in a grammatically correct sentence, they were identifying some concepts in my question as keywords, scanning my sentences within the last few minutes for other concepts with high statistical correlation with these keywords, and then throwing the outcome back at me in a rather unordered form: a rather poorly packaged piece of Artificial Intelligence.

It was a strange experience to witness my students as the embodied proof of the hypothesis of cognitive reductionism that “thinking is a form of computation”. Stranger, though, was the question why all of a sudden half a century after the prime years of cybernetic reductionism we were seemingly having its central thesis[1] actualised.

When I described the algorithm I deciphered to the students and pointed out that they were behaving exactly as a Turing machine would, they enjoyed the scientific explanation without being annoyed at all by the comparison to a machine. When asked how and when they adopted this habit, their immediate and unanimous answer was “while preparing for the university entrance exam”. Indeed, the nation-wide university entrance exam, where every year around 2 million students harshly compete with each other for opportunities of higher education, constitutes a major turning point in their lives. Most students spend typically the last two high school years almost exclusively preparing for this multiple-choice exam “that is

1 “The cyberneticians' first thesis --that to think is to compute as a certain class of machines do—amounted to analyzing and describing what it is to think, not, as it is commonly supposed, to deciding whether it is possible to conceive of machines that think. The question “Can a machine think?” did not come to the forefront until later, at the beginning of the 1950s, as the research program known as artificial intelligence began gradually to establish itself within computer science. To this question cybernetics obviously could only respond in the affirmative, since it had already defined the activity of thinking as the property of a certain class of machines. It is important to see, however, that cybernetics represented not the anthropomorphization of the machine but rather the mechanization of the human.” (Dupuy, 1994, p. 4-59).

taken under severe time pressure. So, it is self-evident that the whole process leaves strong imprints on their studying and answering habits. The combination of high competition, severe time pressure and multiple-choice form forces the students to by-pass the semantics wherever possible and to resort to simple and fast tricks of finding the valid answer.

Nevertheless, this exam existed in this form since several decades and therefore was not sufficient to explain the sudden onset of the new behaviour in that specific academic year. Seeking for better explanations I consulted some of my graduate students who were three or four years elder than the class under consideration. One of them recognised in the behaviour I described a trait, which could have been acquired by playing advanced computer games. As he told, the winning strategies in such games were typically based on identifying the underlying algorithm instead of being "mislead" by the story. So, the hypothesis was that the sophomore class signalled a new generation that grew up playing intensively such computer games, and now they were applying a similar strategy in order to identify a possible correlation between my questions and acceptable answers. Although computer games were most probably not the only cause, this hypothesis, which pointed at the possible effects of "cybernetic encounters" at early ages, was insight providing.

In the subsequent years the students' tendency to miss the essence of education and to seek simple algorithmic routes to success continued. This led to a general decline in their conceptual comprehension. In 2014 in graduate admittance interviews and PhD qualifier exams, where attempts of algorithmic answer generation would be too risky, I have observed students, when asked to explain a concept, providing only the algorithmic description of an operation involving this concept. Retrospectively, I recognise that the sophomore class of 2007 or 2008 was the precursor of a generation that found a safe haven in imitating machine intelligence, which brought with it submission to externally set targets, strong dependence on external appreciation, insufficient self-confidence, and rapid loss of motivation under failure.

Trying to explain the timing of the onset of this behavioural complex, I contrived to consider the socio-political spirit of the years that must have shaped the habits of the sophomore class of 2007-08. They must have been born about mid 1980s and hence must have spent their preschool and early primary school years under the policies of 1990s. But what was so special about these policies?

2 Social Policies and Cybernetics

A brief investigation reveals that rather uniform social policies have been applied in 1990s throughout the Western Bloc (including my country), perhaps only with a few years difference. A vivid illustration of the British case is given in the BBC documentary series by Adam Curtis (Curtis, 2007). As shown in its 2nd episode titled "The Lonely Robot"[2], first the Conservative government (from 1991 onwards) and then the New Labour government (from 1997 onwards) introduced a new system of public management driven by targets and numbers. The system was presented as liberation from the outdated bureaucratic rules, because it gave public servants performance targets and set them free to achieve them in

2 "This episode tells the story of how, in the 1990s, politicians from both the right and the left, tried to extend an idea of freedom, modelled on the freedom of the market to all other areas of society. This was something that previously no-one, not even the high-priest of capitalism, Adam Smith, had thought possible or appropriate. But now, it was seen as inevitable, because underlying it was a scientific model of ourselves, as simplified robots. Rational, calculating beings, whose behaviour and even feelings, could be analysed and managed by numbers." (Curtis, 2007)

any way they wanted. Administrations had no longer the claim of following ideal principles to serve the “public good”. Instead, they were imitating the functioning of the free market, which in return operated on basis of the “negative feedback” principle justified by the science of cybernetics. However, this performance-based approach combined with highly set targets can paradoxically result in the deterioration of the quality (which, however, remains invisible to the evaluation system). As shown in (Curtis, 2007), Episode 2, in some cases public servants who were put under high performance pressure have abandoned the true essence of the tasks and found ingenious ways of hitting their targets. Bizarre examples include hospital managers, who – given the target to reduce the number of patients waiting on trolleys – simply took the wheels off the trolleys, and reclassified them as beds; or police stations, where hundreds of crimes (including assaults, robbery, and fire-raising) were reclassified as simply "suspicious occurrences" in order to have reduced crime rate figures.

In the same period similar social policies have been applied also in my country, and the pragmatic success-oriented mentality underlying these policies has been glorified as “contemporary” and “scientific”, and ever since, performance-based control has been entering almost all domains of life. With the turn of the century the practice of accreditation, which is an implementation of performance-based control, entered my field of vision, because –probably with a few years lag behind West European universities and at least with 10 years lag behind both international and Turkish businesses— my university decided to take an international academic accreditation. During the past one-and-a-half decade the complexity of the accreditation system increased progressively, allowing me to see what a large portion of resources can be stolen from essential tasks only to be wasted on “playing to the gallery”. Now I recognise that the bizarre examples in (Curtis, 2007) bespeak the helplessness of the employees and their loss of confidence in the meaning of the management system rather than their lack of morality.

I will try to describe the traditional and new management approaches in a control theoretical terminology in the hope of identifying why negative feedback approach --known for its efficiency in technology—can fail in the social domain. While attempting this task I will slightly modify the usual control engineering terminology in order to be able to account for some human aspects not applicable in case of technological systems.

3 Within the Control Loop

From the perspective of control theory, “to control a system” means to make it behave in a desired manner by applying appropriate stimuli. Let us call the system, which tries to achieve this, the ***management system***, and break it down into a ***planner*** and a ***controller***. The ***planner*** –based on a model of the “world”-- sets some ***target performance*** that specifies the desired system behaviour, while the ***controller*** is responsible for computing and applying the stimulus (the so-called ***control input***) that will make the system exhibit a performance as close as possible to the target. An open-loop controller (Figure 1.a) can accomplish this control task only under ideal conditions, i.e. in the absence of modelling errors and external perturbations. On the other hand, in the closed-loop scheme (Figure 1.b), where a sensor measures the ***actual performance*** and feeds it back to the ***controller***, these imperfections can be compensated. In the figure a bold arrow has been used for the ***system state*** as opposed to the thin arrow of the ***actual performance*** in order to emphasise that the performance criteria constitute only a small (and discrete) subset of the actual state of the system.

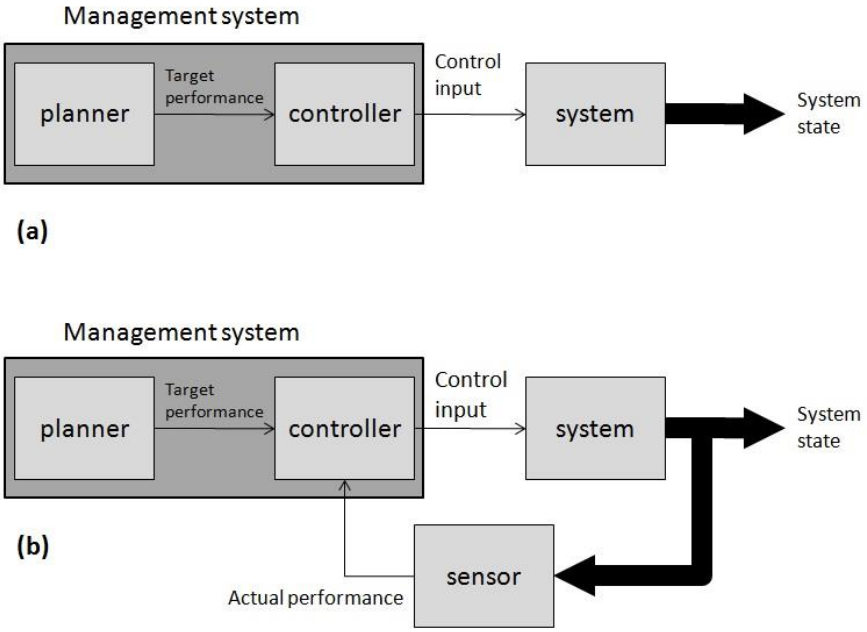


Figure 1: Open-loop (a) and closed-loop (b) control schemes

For the sake of better analysis let us break down the controller in the closed-loop control scheme (Figure 1.b) into a ***comparator***, which calculates the performance error, an ***error correction strategy*** and a ***system model*** (Figure 2). It is worth noting that, given a fixed error correction strategy, the success of the closed-loop controller still depends on the correctness of the ***system model*** within it. For example; a parent, who wants to make a stubborn 3-year old behave in a specific manner, may have to tell the child to do the exact opposite, which he can decide only on basis of a correct model of the child's temper. In order to alleviate the dependence on the correctness of a fixed system model one can incorporate a learning scheme into the controller such that the ***system model*** within the controller can be updated using previously applied ***control input*** and the ***actual performance*** fed back by the sensor (shown by the dashed lines in Figure 2).

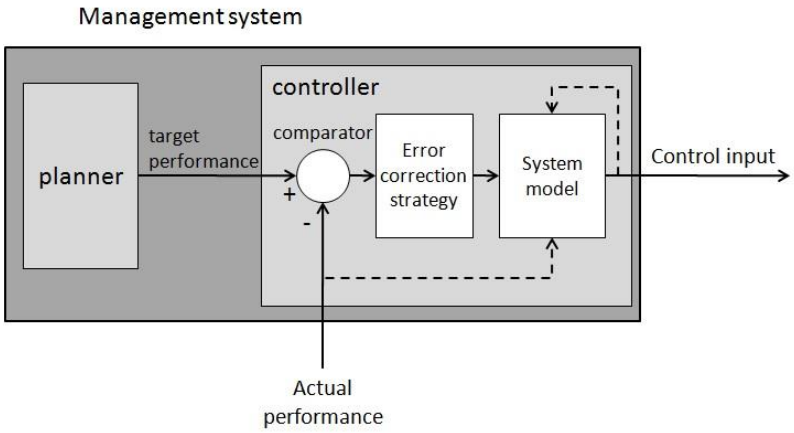


Figure 2: Components of the controller in the closed-loop scheme. Dashed lines indicate the inputs by means of which the system model can be estimated and updated.

In technological applications such a feedback control scheme with eventual adaptation of the system model is proven to provide near perfect control. But when it comes to the control of a learning agent, and more specifically of a human agent, another factor needs to be taken into account: the system has its own ***internal management***. One can envisage this

internal management system as utilising a ***model of the "self"*** as well as a ***model of the "world"*** in order to allocate the agent's resources to various tasks, set goals, evaluate the performance and apply sanctions upon the agent in the form of motivation or self-discouragement. It is also worth noting that the internal management system operates to a great extent at subconscious level.

An attempt to externally manage a system that has its own internal management system bears risks similar to giving a cardiac massage to a person with an intact heart. It is also crucial that the ***world-*** and ***self-models*** of the human agent are adapted and updated (just like the system model within the controller) on experiential basis. Thus, one can say that there exists a competition of learning/adaptation between the ***external management system*** and the human agent. The management system tries to model the human agent, while the latter tries to model the "world", which includes the management system. When the ***external management system*** is less complex than the human agent to be controlled – which is usually the case, as opposed to the scenario with the 3-year old child and the parent-- the external management system cannot win in the modelling competition; but ironically it does not have to, either. Obtaining an accurate ***system model*** is rarely the priority of a ***management system***, which is usually contented with a ***system model*** that provides a feasible trade-off between accuracy and complexity. On the other hand, for any living system modelling the environment correctly (good adaptation of the ***world-model***) and complying with external conditions (good adaptation of the ***self-model***) are evolutionarily advantageous features. As a result of this asymmetry, the interaction of a management system and a human agent can lead to a strange win-win solution, where the human agent wins in the "adaptation game" and complies with the external conditions (as imposed by the external management), while the management system wins in the "control game" because its (usually rather simplistic) ***system model*** has been actualised by the human agent.. Although this solution may appear as a success story from the control point of view, it usually leaves behind a disenchanted and depressed human being who has renounced his true creative capacity and accepted to reduce himself to an automaton-like state.

Fortunately though, whether an attempt of external management of a human agent results in such a tragic compromise, depends on other factors which determine how much the ***world-*** and ***self-models*** of the human agent can resist adaptation. I find three such factors worth indicating here:

3.1 Initial models and limited adaptability

An important factor is related to the fact that the closest social environment plays a decisive role in the formation of the models of a young person, and that usually the ***world-*** and ***self-models*** can be modified only to a limited extent at later stages. Such a limited adaptability can sometimes constitute an advantage that protects the human agent from excessive compliance with the management, if the ***system model*** imposed by the management happens to be in conflict with the initially adopted models. I suppose that this factor can partially explain why the behavioural pattern observed in my sophomore class of 2007 (or 2008) did not appear in earlier generations whose ***world-*** and ***self-models*** were shaped by a social context that was not yet assimilated with new social policies and the underlying mentality. However, one should note that, if the ***system model*** imposed by the external management is in agreement with the agent's initially acquired ***world-*** and ***self-models***, there remains little chance of modification.

3.2 Diversity of influences

The second factor is related to the degree of agreement between various influences on the human agent. (As a matter of fact, the previous item can be considered as a special case of this one.) When the various influences on the human agent balance out each other small personal preferences can make a difference in shaping the internal models. Ironically, the degree of agreement between various influences adversely affects the degree of freedom of the agent. To refer back to my students: this certain sophomore class seems to mark a generation, the various influences on which (the competitive exam system, parents, teachers, peers and playmates --which were literally automata—) reached a critical level of agreement such that I could not help noticing the resulting constructive interference.

3.3 Impact of the management scheme

A third factor is related to the employed management scheme. Different management approaches are not equivalent in terms of the adaptation they induce in the human agent. In order to analyse what has changed in the 1990s, let us try to compare the classical and the performance-based scheme (Table 1).

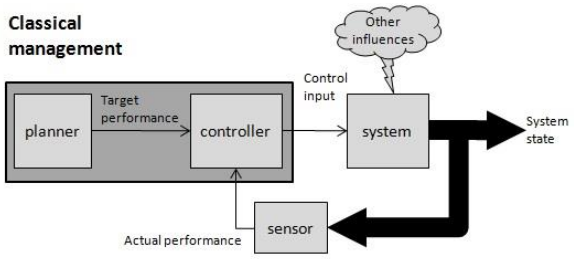
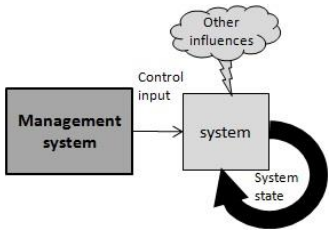
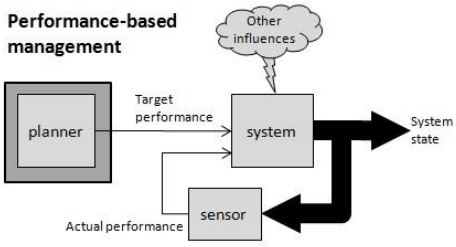
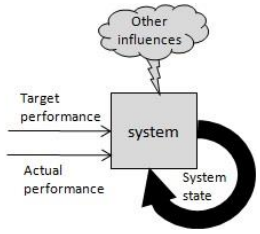
Management scheme	Inputs as perceived by the human agent
	
	

Table 1: Comparison of the inputs as perceived by the human agent in the classical and the performance-based management schemes.

In the classical scheme, in addition to the influences from other sources and the internal observation of the own state, the human agent is directly exposed to the ***control input*** (stimulus) applied by the management system, Depending on the nature and competence of the management system, the character of the ***control input*** can vary from an ingenious guidance to an annoying enforcement. But in any case, the human agent is aware of the fact that it originates from ***external management system***, thus he can take a position in face of it, build a model of the management system, evaluate it, decide whether and to what extent to comply and cooperate with it and set his own targets accordingly.

On the other hand, let us try to model the performance-based management scheme which was introduced by management consultants in 1990s as liberation from the old-fashioned bureaucratic control approach by delegating the ***control task*** to the human agent,

while keeping only the ***planning task*** on the management side. One could regard it as the acknowledgement of the self-management ability of the human agent, a move away from reducing him to an automaton. But actually it only acknowledges the human agent's ability to pursue given targets, while depriving him of the ability to set his own goals and criteria, and to evaluate oneself in terms of them. In this management scheme, in addition to the influences from other sources and the internal observation of the own state, the human agent is directly exposed to the ***target performance*** and the ***actual performance***. But usually in such management applications the source of the ***target performance*** is not sufficiently transparent to the human agent (at least not as transparent as the management system is in the classical management system). In most cases an employee may not know who the planners are, what makes them eligible for this task, and on which basis they decide for the performance criteria and targets. This gives the targets an anonymous character in the eyes of the human agent. Similarly, the ***actual performance*** fed back by a sensor, i.e. a device, appears as an "objective measure" of performance, obscuring the fact that the choice of the performance criteria is a subjective decision on behalf of the planners. Thanks to this alleged objectivity the external performance assessments can push the immediate internal observation of the own state aside, and can enter a dangerous interaction with the subconscious levels of the internal evaluation system, trigger feelings of shame, guilt and inferiority and allow the external evaluation criteria to be internalised. The danger is even larger if –as frequently done-- superiors' performances are indexed to those of their subordinates such that the hierarchical pyramid is both reinforced and used against each member at every level of the hierarchy. Consequently, systematic exposure to such a management scheme can leave permanent traces particularly on the ***self-model*** of the human agent, reducing it to a much simpler one that can be described in terms of the performance criteria defined by the management system. Unfortunately, this description matches too well a symptom I frequently observe in problematic students whose number is growing consistently. These students exhibit an inability to evaluate their own performances independent of external measurements. Although negative outcomes affect them very badly, they cannot live without external measurements because these constitute at the same time the only source of potential appreciation; a symptom I call "score addiction".

Taking also these three factors into account we recognise that, as long as the ***world-*** and ***self-models*** of the human agent correctly captures the essence of the tasks and of the human nature, performance-based control can operate adequately; that is to say, the agent can use the performance criteria and their assessed values as useful clues and indicators for tuning his actions without turning them into ends-in-themselves. So, the adequacy of the ***world-*** and ***self-models*** of the human agent is the guarantor of the operability of the control approach. The deficiency of the performance-based control (when applied to human agents) lies in the fact that it bears the risk of impairing the very factor that guarantees its successful operation. The potential assimilation and deformation of the internal models of the human agent constitutes an additional feedback effect that vitiates the regulatory character of the negative feedback scheme inherent to the performance-based control scheme.

A closer look at the picture reveals another aspect of the difference between technological and social applications of such a control scheme: the true performance of a human agent –as opposed to a machine (no matter how complex)-- cannot be reduced to a finite (no matter how large) number of assessable performance criteria. This is the reason why measured performance indices are just as clues for estimating the true performance

and why it has to be secured that they are considered as such. On the other hand, situations comparable to “taking the wheels off the trolleys and reclassifying them as beds” or “reclassifying serious crimes as simply suspicious occurrences in order to have reduced crime rate figures” (Curtis, 2007) can also occur when the performance-based management scheme in Table 1 is applied to a complex technical system (for the sake of fair comparison let us consider a complex system with its own internal management system). Machine intelligence can similarly find shortcuts to achieve the given targets while giving rise to some undesirable results which are not included in the performance criteria, but this simply means that the planner should include additional criteria such that all relevant aspects of the system output are covered. Unfortunately, that is exactly what most performance-based management applications also do in the social domain. The result is an increase in the bureaucratic load put upon the human agents, rendering them even more prone to “cheating” behaviour. This positive feedback relation leads to the progressive complexification of, for example accreditation systems, as I am personally witnessing nowadays.

Stepping back we can identify further feedback loops in the larger picture: the factors mentioned in 3.1, 3.2 and 3.3 indicate the positive feedback relations between internal models of individuals, the value system of the society and technological paradigms of the era. Once the mentality that considers performance criteria as ends-in-themselves reaches a critical level of prevalence, “the load will shift” in an irreversible manner, and next generations, next policies and next technological paradigms will be shaped by the same mentality.

4 What to change?

No matter how dark the picture is, it harbours a consolation: the socio-political mode of operation that reduces the human to an automaton is not sustainable. In the long (perhaps not so long) run it will lead to system collapse, because the civilised socio-political (and technological) system cannot be kept alive merely by performing well-defined routine tasks. The large system’s survival strongly depends on creativity, something that can be afforded only by human agents in their unreduced state. It asks for an environment where the human potential can flourish in an unlimited and undefined manner. Companies’ desperate investments in “creativity seminars” indicate that businesses are aware of the need while a recent article in *The Guardian* shed light on the academic and scientific side of the picture:

> *“Peter Higgs, the British physicist who gave his name to the Higgs boson, believes no university would employ him in today's academic system because he would not be considered "productive" enough. ... He doubts a similar breakthrough could be achieved in today's academic culture, because of the expectations on academics to collaborate and keep churning out papers. He said: "It's difficult to imagine how I would ever have enough peace and quiet in the present sort of climate to do what I did in 1964." (Aitkenhead, 2013)*

Of course, the prospect that the present trends will lead to a system collapse is not a very relieving argument. One hopes to find a way to break the vicious circle before it leads to a catastrophic end. Here, it is the managements --from businesses, to governments, to supranational institutions-- that have to recognise that our sole resource is the indefinable

human potential, and that its maintenance and furthering is the primary target. Once they acknowledge this, they have to offer incentives in the form of adequate management policies. And since in most cases “the load has been shifted” considerably towards a reductionist conception of the human, the managements will have to do far better than merely preserving “good” internal models (which could have been achieved via a classical management scheme). Now they have to invent open-ended incentives that will allow the human to rediscover his true potential.

These expectations may seem paradoxical in view of the fact that managements consist of human agents who are in the same boat and subject to the same shift of values. Nevertheless, the strong negative feedback of a threat of system collapse may be mind opening. Let us hope that managements recognise as their primary goal the maintenance and furthering of the potential, the freedom, the self-respect and self-confidence of the human, before more generations fell prey to the mechanisation of the mind.

References

Dupuy, J-P. (2009). *On the Origins of Cognitive Science: The Mechanization of the Mind*. Cambridge: The MIT Press.

Curtis, A. (2007). *The Trap: What Happened to our Dreams of Freedom*. BBC documentary film.

Aitkenhead, D. (2013). Peter Higgs: I wouldn't be productive enough for today's academic system. *The Guardian*, Friday 6 December 2013.

About the Author

Yagmur Denizhan

Professor at the Electrical and Electronics Engineering Department and Head of the Graduate Program in Systems and Control Engineering of Bogazici University. Recent technical research areas are nonlinear dynamics, chaos control and modelling of biological systems, while wider fields of philosophical involvement include cognitive science, biosemiotics, systems theory and mythology.