Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

# Probability of Random Execution of the Generated Reaction as a Hidden Parameter of Forming Instrumental Reflexes

**Alexander B. Saltykov[1], Sergey V. Grachev[2], Sergey B. Bolevich[3]**

[1]I.M. Sechenov Moscow State Medical University, Moscow, Russia. E-mail: absaltykov@yandex.ru

[2]I.M. Sechenov Moscow State Medical University, Moscow, Russia.

[3]I.M. Sechenov Moscow State Medical University, Moscow, Russia.

**ABSTRACT**

The a priory probability of random correct reaction (PRCR) constitutes a hidden parameter of instrumental (operant) learning. An algorithm for computer modelling has been developed enabling assessment of the PRCR effect on the learning process in a probabilistically organised environment. No consideration of empirical biological data is required in the algorithm; they can be used only to verify (test) the identified regularities. Good agreement between the data from the literature and the results of biological experiments conducted by the authors indicate that PRCR should be taken into account, when identifying the zones of optima and pessima in the space of parameters that influence learning. The obtained results can be used in planning biological experiments, and reproducing instrumental reflexes in domesticated social robots, autonomous reconnaissance drones, autonomous car control systems, digital characters of computer games.

**KEYWORDS**

Instrumental Reflex, Random Correct Reaction, Probability of Random Reaction, Computer Modelling, Artificial Intelligence.

## Introduction

Enhancing artificial intelligence implies the development of new algorithms for enabling technical devices to autonomously assess the essence of events, for which they were not initially designed and, at the same time, associated with achieving the purpose of activity. Throughout biological evolution, this problem had been solved using conditioned reflexes, allowing to reveal a signalling meaning ("essence") of conditioned stimuli, the occurrence of and dynamic changes in which cannot be foreseen and programmed at the level of genotype. Conditioned reflexes as a form of anticipation facilitate predicting and prior preparation for forthcoming events (Skinner, 1938; Lorenz, 1981; Saltykov, Grachev, 2015). The adaptive significance of this activity implies its consideration in the works devoted to artificial intelligence.

Instrumental conditioned reflexes ensure implementation of a certain reaction (usually, motor) in connection with the conditioned signal to further receive the reinforcement. They are formed based on a dynamic stereotype that reflects probabilistic and statistical regularities. Even B.F. Skinner (1938) and I.P. Pavlov (1927) were writing that in addition to the conditioned signal there are multiple other stimuli, antecedent to the reinforcement: random sounds, convective air motions, the researcher's movements, etc. Hence, to distinguish such a signal from many other stimuli, at least several combinations of conditioned and unconditioned stimuli are usually needed (Balsam P.D., 1985; Burgos J.E., 2019; Simonov P.V., 1991). On the contrary, instantaneous initiation of reflex in response to any influence prior to reinforcement, is biologically inexpedient; imprinting (Lorenz K., 1981) and some other reflexes with reinforcement of extreme importance for an individuum constitute rare exceptions. Even in "strictly determined" conditions with inherent unique dependence of the conditioned signal and reinforcement, the first signs of learning usually appear not earlier than 4-5 combinations occur; extinction of reflexes takes place at the same rate. This dependence does not depend on the level of phylogenetic development: the experiments were made on earthworms, grapevine snails, fish, bees, rats, cats, dogs, and other animals, and humans as well (Skinner B.F., 1935; Batuev A.S., 2005; Honig W.K., Staddon J.E.R., 1977; Staddon J.E., 2016).

There is also a probabilistic nature of relationship between the conditioned signal and reinforcement in mammals and birds. It has been first studied in the I.P. Pavlov experiments related to differentiation of similar stimuli, one of which was conditioned (oval and circle, audio signals similar in frequency, etc.) (Pavlov I.P., 1932). A probabilistic mode of reinforcing generated reactions, near-threshold intensities of the conditioned stimulus, variation of interstimulus intervals have also been extensively used. Here, low rate of training, frequent neurotic disorders and states of learned helplessness, and a considerable statistical scatter in the acquired data are inherent (Amsel, A. 1967; Davison V.,

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

Jenkins P.E., 1985; Jenkins P.E., 1985; Domjan M., 2014; Dwyer D.M. et al., 2019). Conducting experiments and their comparative analysis are further complicated by the fact that there are multiple parameters of probabilistic learning environment, the values of which may substantially vary. This all explains, why studying of the respective regularities has not yet been completed, and most studies have thus far been conducted in "strictly determined" conditions.

The a priori probability of random correct reaction (PRCR) is the least-studied parameter of instrumental learning, usually not even taken into account (!) in the context of conventional biological and psychological concepts (Skinner, 1938; Commons, 1991; Honig, 1977; Staddon J.E., 2016). However, the value of this "hidden" parameter is responsible for the number of initially random search reactions, which are executed due to the conditioned signal and obtain the respective reinforcement. Owing to this, PRCR modulates the information interaction with the environment during orientational and search activity. In "strictly determined" conditions, the modulating effect is minor, however, it can be very significant, when there is a complex relationship between the conditioned signal and reinforcement (Saltykov A.B., 2013; Saltykov A.B. et al., 1990). Unfortunately, a considerable amount of effort required for biological researches in a probabilistically organised environment impedes studies of this parameter. Scarce empirical data, in their turn, hinder the development of simple and effective mathematical models for technical devices.

The computer modelling of the process of formation and extinction of instrumental reflexes may facilitate solving the problem. It explicitly accommodates the effect of PRCR and other key parameters: probabilities of receiving reinforcement in connection with the conditioned signal and with no connection to its presentation, and statistical significance of the decisions made. The Monte-Carlo method does not require quantitative description of relationships between variables based on the pre-acquired empirical data. Hence, correspondence between the results of modelling and the data already available in the literature will indicate that the utilised approach is heuristically valuable. On the one hand, it will predetermine relevance of its using in technical devices, and, on the other hand, for predicting still unknown regularities of training bio-objects in a probabilistically organised environment.
The work was aimed at developing the algorithm for computer modelling of instrumental (Skinner) reflexes in a probabilistically organised environment, which takes the PRCR effect into account. Here, "strictly determined" conditions were considered as a particular case.

## Materials and Methods

### The Assumptions Used

(1) The process of training technical devices and living organisms implies identical information provision. It enables developing computer algorithms based on general biological concepts and, vice versa, defining biological data more precisely using computer modelling.
(2) The information, needed to form the instrumental reflex, is acquired due to orientational and search (test) instrumental reactions. Some of such reactions are executed randomly in connection with presentation of the conditioned reflex and receive the appropriate reinforcement (PRCR > 0).
(3) Random nature of orientational and search activity implies that instrumental reactions are distributed uniformly over time, on average, in relation to the conditioned stimulus. It enables theoretical assessment of the PRCR value as the ratio between the total duration of all presentations of the conditioned signal and the total duration of the experiment.
(4) After each orientational and search reaction, a probabilistic decision is made on the availability or absence of relationship between the conditioned stimulus and reinforcement. The similar decision is made with respect to any other stimulus, antecedent to the unconditioned reinforcement.
(5) If the probabilistic decision is made on the availability of relationship between the conditioned stimulus and reinforcement, search instrumental reactions cease being random, i.e. the probability of their correct execution (in connection with the conditioned signal effect) drastically increases. The number of search instrumental reactions ($N$), required for intermittent transition from the untrained state to the trained one, is obtained as a result of modelling. It shall be emphasised that this assumption is a matter of dispute, since the process of forming the trained state can be also described by the S-shaped curve (Skinner B.F., 1938; Honig W.K., Staddon J.E.R., 1977).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

(6) Computer modelling of making probabilistic decisions is possible based on statistical criteria. The level of significance (the 1st type error, α) demonstrates inertia in the operation of corresponding neurophysiological mechanisms (risk-taking in decision-making) and formalises predisposition of a learner to the reflex formation. The more "capable" and motivated a hypothetical learner is, the higher his genetic predisposition to the generated reflex is, and so on, the less number of search reactions he will need on average for identification of the existing regularities and transition to the trained state: setting α at a relatively high level is a formal reflection of it. On the contrary, when biological importance of reinforcement is insufficient or certain segments of brain are damaged, even a great number of search reactions will not enable a learner to reveal the relationship between the signal stimulus and reinforcement: low level of α should be established, since the right decision can be made only in "obvious" cases, when the probability of error is close to zero.

**Methodology of Computer Simulation**

Test instrumental reactions, which permit to identify the relationship between the conditioned signal and reinforcement, can occur at any moment of time. Hence, probability *p(a)* of their random execution within the intervals, during which the conditioned signal occurs, is the ratio between the duration of these intervals and the total duration of the experiment: *in computer modelling, it was found that presentation of the PRCR value in the p(a) form is more convenient*. In this case, erroneous performance of the generated instrumental action, not associated with presenting the conditioned signal, has the probability of *1 – p(a)*. In addition, in the "random" environment the *p(κ/a)* probability of receiving reinforcement in connection with the conditioned signal effect can be less than 1, and with no connection to this signal – *p(κ/b)* - is more than zero. These parameters together with the α indicator (the 1st type error) are sufficient for computer simulation.

Suppose that the reflex of lever-pressing is modelled during the effect of the conditioned stimulus using positive (food) reinforcement. Let *N* pressings already be made in the course of orientational and search activity. Then, the probabilistic average of the number of lever-pressings, randomly occurred during the period of the conditioned signal effect, is *N. p(a)*, and the number of the received therewith positive reinforcements - *N. p(a). p(k/a)*. Computation for the other cases presented in the matrix of possible outcomes (**Table**) is made in a similar way; when filling the matrix, the absence of positive reinforcement is considered negative. If necessary, the same matrix can reflect the use of negative stimulation, for example, electrodermal stimulation (its absence will be considered as positive reinforcement) – such an "inversion" corresponds to biological concepts, despite known differences between reactions to various types of reinforcement (Schindler C.W., Weiss S.J., 1982; Gershman SJ., 2015).

**Table.** Probabilistic average of the number of reinforcements of various modalities, obtained during learning process

| **Conditioned signal** | **Number of reinforcements** | |
|---|---|---|
| | **positive** | **negative** |
| presented | $N.\ p\ (a).\ p\ (k/a)$ | $N.\ p\ (a).\ [1 - p\ (k/a)]$ |
| absent | $N.\ [1 - p\ (a)].\ p\ (k/b)$ | $N.\ [1 - p\ (a)].\ [1 - p\ (k/b)]$ |

Learning process itself is simulated as follows. A generator of random numbers, uniformly distributed from zero to 1, distributes each imitated search reaction in one of cells of a 4-field table, taking into account pre-specified learning parameters. It shall be done in two stages: at the first stage, the time of executing the first instrumental reaction is identified (with the effect of the conditioned signal or without it), and at the second stage, the modality of the received reinforcement is found (absence of positive reinforcement is considered negative). After completing the second stage, 1 is added to the content of the respective cell.

Suppose that the authors are concerned with the number of test lever-pressings (*N*), necessary for forming the reflex in the following conditions: *p(a)* = 0.4; *p(k/a)* = 0.9; *p(k/b)* = 0.3. The first pressing shall be simulated using the random-number generator: if the number is less or equal to 0.4, pressing is considered performed against a background of the conditioned signal effect; otherwise – with no connection to it. Then, it shall be defined, whether the positive reinforcement was received: to this end, the next random number is generated (to be definite, it shall be assumed to be 0.8). Then, if the instrumental reaction occurred within the time of the conditioned signal effect, the reinforcement is considered received [0.8 < *p(k/a)*]. If lever-pressing was executed with no connection to the

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

conditioned stimulus, there is no positive reinforcement [$0.8 > p(k/b)$]. 1 is added to the content of the respective cell, depending on the obtained result.

After simulating each instrumental reaction, according to the $\chi^2$ criterion for the 4-field table one of the two statistical hypotheses is chosen: $H_0$ – there is no relationship between presenting the conditioned signal and reinforcement and $H_1$ – there is a relationship between presenting the conditioned signal and reinforcement. In case of accepting the zero hypothesis, the next instrumental reaction is "generated". On the contrary, selecting the alternative hypothesis means formation of the reflex, since the subsequent instrumental reactions will be executed only in connection with the conditioned signal. The $\chi^2$ statistical criterion is used to simulate the mechanism of making probabilistic decisions, and the 1st type error (α) is specified as an independent parameter. Accepting the alternative hypothesis doesn't mean that computer modelling is stopped, it can be continued to assess the rate of extinction of the already formed reflex with corresponding "sudden" decrease in $p(к/a)$, increase in $p(к/b)$, etc. In this case, the matrix of possible outcomes will continue to be filled until the moment of accepting the zero statistical hypothesis, and the rate of reflex extinction is assessed by the number of instrumental reactions required to that effect.

The developed algorithm allows to reproduce formation of reflexes not only with probabilistic mode of reinforcing instrumental reactions. It is also suitable for studying near-threshold intensities of the conditioned signal. Although the intensity of the conditioned signal is not an individual variable, its influence can be examined through combining the values of other parameters. It follows from psychophysical concepts of the probabilistic nature of perceiving near-threshold stimuli (Green D.M., Swets J.A., 1966; Schonhoff T.A. et al., 2006; Clifton R.K. et al., 1994). Suppose, there is a need to model the process of generating 100% reinforced reflex in response to a weak sound stimulus, correctly perceived by a bio-object with 0.85 probability. It means that 85% on average of unconditioned reinforcements will be associated with the conditioned stimulus, and 15% - with its absence: $p(k/a) = 0.85$; $p(к/b) = 0.15$. Modelling the process of differentiation of similar stimuli, one of which is conditioned, is analogous (oval and circle, audio signals similar in frequency, etc.)

It has earlier been emphasised that usually there are multiple random (not related to reinforcement) stimuli, besides the conditioned stimulus. In computer modelling, an individual matrix of possible outcomes should be formed for each such stimulus. The statistical analysis of such matrices will not allow to identify statistically significant relationships. An independent analysis of each stimulus is in agreement with the concept of cognitive strategies in situations of multi-alternative choice (von Helversen B., 2013; Moreno-Rios S., 2014). It is considered that when solving indeterminate multiple-model problems the use of simple problems is preferable (Ragni M., Knauff M., 2013), despite an incomplete (representation) of empirical data (Chevalley T., 2016).

Thus, computer modelling enables predicting the number of search instrumental reactions (*N*), needed for formation and extinction of instrumental reflexes. It can be performed with different combinations of the PRCR values and other parameters of a probabilistically organised environment. Making probabilistic decisions (on existence or absence of relationship between the conditioned signal and reinforcement) is modelled using a statistical criterion. The value of the 1st type error (α) is a single parameter of modelling, which characterises risk-taking in decision-making. It reflects inertia in the operation of decision-making mechanisms, their "predisposition" to the reflex formation.

## Results and Discussion

50 simulation experiments were performed for each studied combination of learning parameters, what allowed to compute an arithmetic average and confidential interval *N*. The results were compared with the data of biological experiments. It has been found that the predicted value of *N* tends to infinity in case when probabilities of reinforcing instrumental reactions, executed in connection and with no connection to the conditioned signal [$p(k/a) = p(к/b)$] are equal. It corresponds to biological concepts (Skinner B.F., 1938; Simonov P.V., 1991; Staddon J.E., 2016), since there is no adaptive significance in the generated reflex.

Most findings were presented in the form of three-dimensional graphs (examples are given in Fig. 1, 2), visually demonstrating the value of the $p(a)$ parameter. The graphs show a growing complexity of learning (an increase in *N*), as the $p(a)$ parameter approaches zero. It corresponds to the known problematical nature of training bio-objects when using unduly short-term conditioned stimulation, what is traditionally explained by the problems of perception

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

(Green D.M. et al., 1966; Schonhoff T.A. et al., 2006). But, if the intensity of such stimulation a priori exceeds a threshold of perception, an unduly small value of *p(a)*, i.e. almost full impossibility of random execution of search reactions in connection with the conditioned signal, should, in the authors' opinion, take on crucial importance. Zero duration of the conditioned signal implies obvious impossibility of forming reflex, what corresponds to *N*, infinite at *p(a)* = 0.

It is known that too long-term presentation of the conditioned signal is also not good for learning due to delay in unconditioned reinforcement (Richards R.W., 1986; Staddon, 2016). In modelling, it is reproduced by the progressive increase in *N* as the *p(a)* parameter approaches its maximum possible value. When *p(a)* = 1, impossibility of forming reflex is foreseen; in biological studies this situation can be reproduced by continuous presentation of the conditioned stimulus within the entire period of "learning" – only in this case all the instrumental reactions will be formally correct, i.e. executed in connection with the conditioned signal.

Fig.1 demonstrates the maximum rate of learning in the "strictly" determined environment [*p(k/a)* = 1.0; p(к/b) = 0]. But, even in this case, it is supposed that there are at least several combinations of conditioned and unconditioned stimuli. It once again is in agreement with the data of biological experiments: at least 4-5 combinations are usually required for learning (Skinner, 1938; Burgos J.E., 2019). However, if the α parameter (the 1st type error) has rather high value, the modelled reflex can be formed even after the first combination. Such high predisposition to learning in bio-objects is observed in imprinting (Lorenz K., 1981), and when applying extremely significant reinforcing stimulation (Batuev, 2005).

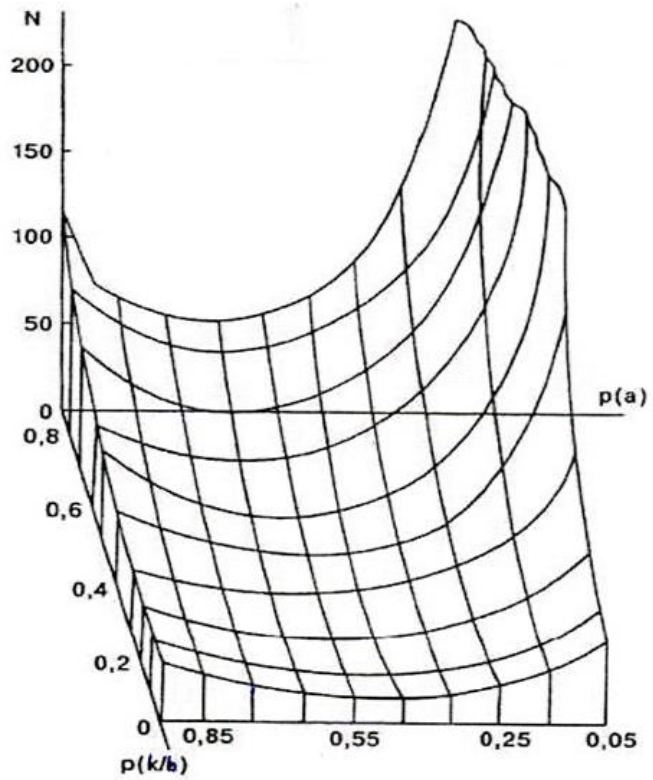


**Fig. 1.** Relationship between the rate of learning and parameters of "random" environment at the 100% probability of reinforcing correct reactions [*p(k/a)* = 1.0]. Along the axis of abscises: *p(a)* of random execution of the generated reaction; along the axis of ordinates: *p(k/b)* of positive reinforcement of incorrect reactions; in the z-direction: the number of search reactions required for establishing relationship between the conditioned signal and reinforcement; $\alpha$ = 0.001- the value of the 1st type error, the achievement of which means identification of relationship between the conditioned signal and reinforcement. Further explanations are given in the text.

It is well known that the probabilistic mode of reinforcing correct instrumental reactions slows down the process of reflex formation (Amsel, A. 1967; Davison V., Jenkins P.E., 1985; Domjan M., 2014; Dwyer D.M. et al., 2019). This

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

effect is also reproduced in computer modelling [$p(k/a) < 1$, $p(\kappa/b) = 0$]; Fig.1 [$p(k/a) = 1.0$] and Fig.2 [$p(k/a) = 0.55$] can be compared as an example, since when constructing them the identical value of the 1st type error was used. Another known effect is also reproduced: relatively slow extinction of reflexes, generated in conditions of probabilistic mode of reinforcing (Kinlstom J.F., 1987; Domjan M., 2014; Bouton M.E., 2019). The formed reflex can become extinct when there is a "sudden" decrease in $p(\kappa/a)$, an increase in $p(\kappa/b)$ and other loss of information significance in the conditioned signal.

Fig. 1 shows that in case of 100% reinforcement of correct reactions, there is almost no effect of minor exceedances of the $p(\kappa/b)$ parameter zero level (probabilities of receiving positive reinforcement with no connection to the conditioned signal) on the rate of learning. A completely different type of situation occurs when $p(\kappa/a)$ – probability of positive reinforcement in connection with the effect of the conditioned stimulus – turns out to be substantially less than 1 (Fig.2). In this case even minor change in the value of $p(k/b)$ throughout its entire variation interval noticeably affects the rate of learning. It has been this, the experiments performed on animals and a human have been indicative of (Grey D.A., 1978; McNamara J.M. e.a., 1983; Miltenberger e.a., 2016). Fig. 1 and 2 also show that the reflex is formed substantially faster as the $p(a)$ value approaches 0.5. The forecasted changes are relatively small with 100% reinforcement of generated reactions (Fig.1) and considerably increase when there is a probabilistic mode of unconditioned reinforcement (Fig.2). This regularity can be considered as a forecast for computer modelling, which requires its justification in biological experiments.

The utilised algorithm of modelling reflects a well-known complexity in forming reflexes when the conditioned stimulation has a near-threshold intensity. Although the intensity of the conditioned stimulus is not an individual parameter of modelling, its effect can be taken into account indirectly. It shall be illustrated by the example of forming 100% reinforced instrumental reflex in response to the near-threshold conditioned stimulus. Suppose, this stimulus is perceived correctly with 0.55 probability. Then, according to the psychophysical theory of detecting the signal (Green D.M., Swets J.A., 1966; Schonhoff T.A. et al., 2006; Clifton R.K. et al.,1994), 55% of unconditioned reinforcements will be associated with presenting conditioned stimulus, and 45% - with its absence: $p(k/a) = 0.55$; $p(\kappa/b) = 0.45$. Computer modelling demonstrates that formation of reflex in such conditions is highly problematic, i.e. high values of $N$ (Fig. 2). The effect of a stronger conditioned signal is simulated through increasing the difference between parameters $p(k/a)$ and $p(\kappa/b)$ – forecasted acceleration in learning also corresponds to the literature data (Cooper L.D. et al., 1990; Clifton R.K., e.a., 1994; Commons M.L., 1991; Macmillan N.A., 2005).

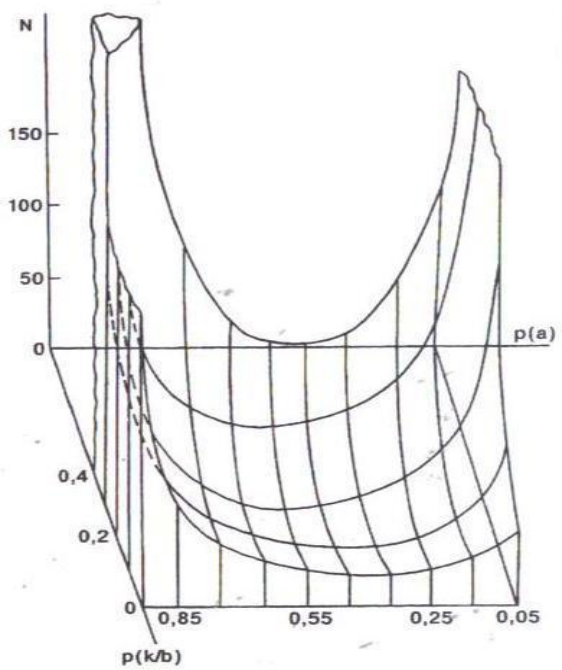


**Fig. 2.** Relationship between the rate of learning and parameters of "random" environment at the 55% probability of reinforcing correct reactions [p(k/a) = 0.55; α = 0.001]. The other designations are the same as in Fig.1.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

Computer modelling reproduces one more known effect: slowing down of the rate of learning, if the conditioned stimulation without respective reinforcement precedes it (Larats D.B. et al., 1988; Bouton M.E., 2019). It shall be illustrated by the specific example. Suppose that the conditioned reflex is formed in the "strictly determined" conditions: $p(k/a) = 1$, $p(\kappa/b) = 0$. However, repeated presentations of the conditioned stimulus without any reinforcement precede the process of its formation – in computer simulation it means filling of cells of only right column of the matrix of possible outcomes (table). Such filling substantially slows down achieving the critical value of the α significance level with subsequent simulation of the conditioned-reflex learning as such.

The most essential result of computer modelling involves the potential examination of the *p(a)* parameter, on which the probability of executing search instrumental reactions in connection with the conditioned stimulus and receiving the respective reinforcement depend. The parameter modulates the information significance of search reactions and, owing to this, determines the location of the zones of optima and pessima in the space of parameters that influence learning (Fig. 1, 2). The forecasted effects are particularly significant in a probabilistically organised environment [$p(k/a) < 1$, $p(\kappa/b) > 0$]. At the same time, *this parameter is traditionally not considered when carrying out biological researches* (Commons, 1991; Skinner, 1938; Honig, 1977; Staddon J.E., 2016). *Moreover, the description of the experiment methodology accepted in the literature does not suffice to accurately compute it*. It makes comparative analysis of biological studies, conducted in different conditions (at various probabilities of random execution of the generated reaction), complicated (Saltykov A.B., 2013; Saltykov A.B. et al., 1986).

In the authors' opinion, the assessment of the *p(a)* value should be based on the fact that before establishing the relationship between the conditioned signal and reinforcement, search instrumental reactions are distributed uniformly over time, on average, in relation to the conditioned stimulus. It enables determining the probabilistic average of the *p(a)* value as the ratio between the total duration of all presentations of the conditioned stimulus and the total duration of the experiment. Such *idealisation is convenient for computer modelling and is based on the assumption that execution of test instrumental reactions is possible at any moment of time*. With that, real biological studies involve rather long-term periods, during which no new search reactions can be executed. These are the intervals, utilised by a bio-object to receive unconditioned reinforcement and to further resume search activity (cessation of chaotic movements, for example, after electrodermal stimulation), and the time spent to execute instrumental reactions as such. Unfortunately, the duration of these time intervals is usually not specified in scientific articles, although to define the value of *p(a)*, they should be excluded from the computation formula. Strictly speaking, only after this the results of biological experiments can be properly compared with the data of computer modelling. A correction is not required only in case when the *p(a)* parameter is equal to zero or 1 – in biological studies it corresponds to uninterrupted presentation of the conditioned signal throughout the entire experiment [$p(a)=1$] or its complete absence [$p(a)=0$], i.e. when learning is basically impossible.

The experiments on animals with precise calculation of the *p(a)* value were conducted in conditions of a probabilistically organised learning environment. The probabilistic mode of unconditioned reinforcement (Saltykov A.B., 2013; Saltykov A.B., Toloknov A.V., Khitrov N.K., 1989, 1990) and near-threshold intensity of the conditioned stimulation were in use (Saltykov A.B., 2013; Saltykov A.B., Toloknov A.V., Khitrov N.K., 1993, 1998). The results of experiments have proved to be in good agreement with the data of computer modelling and confirmed the influence of the *p(a)* parameter on information interaction between an individuum and the environment. Optimal for learning values of this parameter really facilitate the reflex generation when there is a probabilistic mode of unconditioned reinforcement. On the contrary, pessimal values of *p(a)* decrease informational significance of search reactions, especially, in a probabilistically organised environment. It should be emphasised that the obtained results are preliminary, and additional studies are needed. However, it has been even now obvious that it is worthwhile to consider the effect of the *p(a)* parameter on the conditioned-reflex activity.

Hence, the developed algorithm enables effectively predicting the rate of generation and extinction of instrumental reflexes. It doesn't imply the use of empirical data – they can be involved only to verify (test) the obtained results. The correspondence between these results and the data of biological experiments proves the heuristic potential of computer modelling, the key element of which is the *p(a)* parameter. Its influence is particularly significant in a probabilistically organised environment, what should be taken into account when training not only living organisms, but technical devices as well.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

## Instrumental Reflexes in Technical Devices

In this case, search instrumental activity is conducted by the device itself, rather than simulated using the random-number generator. Some search reactions are executed randomly due to the conditioned signal [$p(a) > 0$], and the others – with no connection to it. Each such reaction is accompanied by positive or negative reinforcement. Nowadays, there already exist robots, capable of generating the simplest conditional reflexes (Parker G.B., 2007; Brooks, 2014; Weiss et al., 2015). By analogy with bio-objects, they perceive stimulation of certain sensors as "emotionally" negative or positive reinforcement.

The matrix of possible outcomes (**Table**) is filled by analogy with the computer modelling procedure. When utilising reinforcement of only one modality, absence of positive reinforcement is perceived as negative, and vice versa. After each instrumental reaction, one of the two statistical hypotheses is chosen: $H_0$ – there is no relationship between presenting the conditioned signal and reinforcement and $H_1$ – there is a relationship between presenting the conditioned signal and reinforcement. Accepting the zero hypothesis implies that the search activity will be continued. On the contrary, selecting the alternative hypothesis means that the required relationship is found, and then all the subsequent instrumental reactions will be executed only in connection with the conditioned signal. The value of the 1st type error in the $\chi^2$ criterion is pre-programmed considering the expected value of the generated reflex.

The available formed reflex doesn't mean that analysing the matrix of possible outcomes is ceased – its filling should be continued even after accepting the alternative hypothesis. It will enable monitoring the potential loss of the adaptive value of the already generated reflex, when learning parameters change: $p(k/a)$, $p(k/b)$, $p(a)$. And if a zero statistical hypothesis is accepted at some moment, the previously formed reflex immediately becomes extinct – the technical device will continue the implementation of instrumental reactions already in a random relation to the signal, which lost its information significance.

In addition to the conditioned signal, multiple random (not associated with reinforcement) stimuli can influence the robot sensors. In relation to each of them, an individual matrix of possible outcomes (table) should also be formed, and a statistical analysis should be carried out. And if the analysed stimuli are really not related to reinforcement, the respective instrumental reflexes will not be generated.

It should be emphasised that the formation and extinction of instrumental reflex does not imply that a technical device will compute the $p(a)$ value and other parameters of a probabilistically organised environment. The analysed matrix of possible outcomes accumulates only final results of search instrumental activity, what levels out the effect of various parameters. Computation of the $\chi^2$ criterion does not require pre-determining each of them. Concurrently, a quantitative assessment of parameters, influencing learning, is necessary for *computer modelling of the future events* based on the current orientational and search activity. Such modelling can be useful in the cases of high uncertainty of learning environment: near-threshold intensity of the conditioned signal, low probability of unconditioned reinforcement, a need for distinguishing the conditioned signal from multiple other stimuli, etc. Computer modelling enables predicting the rate of forming the reflex (the $N$ value), which should be taken into account, when making a decision on whether it is expedient to continue learning in the existing conditions. For instance, identification of relationship between the analysed ("conditioned") signal and reinforcement should always be ceased, if the predicted value of $N$ tends to infinity. Such a situation is possible, if the value of the $p(a)$ parameter approaches zero or 1. The identical regularity is observed, when the probabilities of unconditioned reinforcement of search reactions, executed in connection with and with no connection to the analysed signal, are equal [$p(k/a) = p(\kappa/b)$].

An approximate assessment of values $p(a)$, $p(k/a)$, $p(k/b)$, required for computer modelling of the future events, is possible based on the current orientational and search activity. Value $p(k/a)$ is defined as a part of search reactions, actually executed in connection with the analysed signal and received the respective reinforcement. Value $p(k/b)$ is computed as a part of instrumental reactions, executed with no connection to this signal and received the same reinforcement. Value $p(a)$ constitutes the ratio between the total duration of all presentations of the "conditioned" signal and the total duration of the experiment (time intervals, during which no new instrumental reactions can be executed, should not be taken into consideration).

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

Thus, it seems appropriate to take into account the effect of the "hidden" *p(a)* parameter on formation and extinction of instrumental reflexes in technical devices. It is possible within the framework of the developed algorithm, and it allows to optimise the process of learning in a probabilistically organised environment: when the intensity of the conditioned signal is near-threshold, when there is a probabilistic mode of unconditioned reinforcement, and when a need arises for distinguishing the conditioned signal from the noise of stimuli, etc. Not only the simplest audio, light, or mechanical stimuli can be utilised as the conditioned and/or unconditioned stimuli, but also the images of subjects, people, and animals, recognisable by present-day technical devices (Fan B., 2015; Awad A.L., 2016). It widens the sphere of applying the developed algorithm for training domesticated social robots, autonomous reconnaissance drones and submarine devices, autonomous car control systems, and systems for correcting behaviour of computer games' digital characters.

## Conclusion

The effective algorithm for computer modelling of instrumental (operant) learning in a probabilistically organised environment has been developed with its most essential feature of taking into account the effect of the a priori probability of random correct reaction (PRCR), considered in computer modelling as an individual parameter – *p(a)*. This parameter is traditionally ignored in the context of biological and psychological studies (Commons, 1991; Skinner, 1938; Honig, 1977; Staddon J.E., 2016), what complicates its use in mathematical models and works on artificial intelligence (Scozzafava R., 2007; Kim Y.T. et al., 2010; Longbing, 2012; Hinton G.E., 2014; LeCun Y. et al., 2015; Kemp C. et al., 2011; Spektor M.S., Kellen D., 2018). However, just the value of this parameter defines what number of search instrumental reactions in the process of learning will be executed due to the conditioned signal and will obtain the respective reinforcement. And it, in its turn, modulates the information significance of orientative and search activity, especially in a probabilistically organised environment (probabilistic mode of reinforcement, near-threshold intensity of the conditioned stimulus, etc.).

As opposed to the equations of mathematical learning theory (Bower G.H., 1994; Fulop S, Chater N., 2013), no preliminary computation of the equation coefficients is implied in the developed algorithm; the empirical data are utilised only to test the already obtained results. The achieved conformance with the data of biological experiments indicates that the consideration of the *p(a)* parameter facilitates reproduction of the conditioned reflex activity in living organisms. An opportunity arises for efficient prediction of the behaviour dynamics even after the first combinations of the conditioned and unconditioned stimuli, what is challenging in the context of other approaches (Scozzafava R.S., 2007; Calvin N.T. et al., 2015; Burgos, 2019). The conducted research can be considered as an alternate version of the Turing test (Turing A., 1950) on differentiation between the inherent in living organisms and artificially generated forms of activity.

## Disclosure Statement

No potential conflict of interest was reported by the authors.

## Notes on Contributors

***Saltykov Alexander***, M.D., Professor at the human pathology department, Sechenov First Moscow State Medical University (Sechenov University). He integrated the Pyotr K. Anokhin's theory of functional systems with the concept of probabilistic forecasting and substantiated a need for differentiating several backbone factors within the framework of this theory. He is interested in conditioned reflex learning under conditions of subjective uncertainty: probabilistic schedule of conditioned and unconditioned stimuli, near-threshold intensity of signals in use, presence of competing motivations, etc.

***Grachev Sergey***, M.D., Professor at the human pathology department, Sechenov First Moscow State Medical University (Sechenov University), member of the Russian Academy of Sciences. His research interest is pathophysiology of extreme states.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

***Bolevich Sergey***, M.D., Professor, head of the human pathology department, Sechenov First Moscow State Medical University (Sechenov University). He is interested in free radical processes under normal and pathological conditions.

## References

[1] Amsel, A. (1967). *Partial reinforcement effects on vigour and persistence.* In: Spence K.W., Spence J.T. (eds.) The psychology of learning and motivation, New York: Academic Press, 1-65.

[2] Awad, A.L., Massaballah M. (eds.). (2016). *Image feature detectors and descriptions*. Springer, 134.

[3] Balsam, P.D. (1985). *The functions of context in learning and performance.* In: Balasm P.D., Tomie A. (eds.) Context and learning, Hillsday, New Jersey, 1-21.

[4] Batuev, A.S. (2005). *Physiology of high nervous activity and sensory systems*. Saint-Petersburg: Peter, 317.

[5] Bouton, M.E. (2019). Extinction of instrumental (operant) learning: Interference, varieties of context, and mechanisms of contextual control. *Psychopharmacology (Berl)*, 236(1): 7-19.

[6] Bower, G.H. (1994). A turning point in mathematical learning theory. *Psychological Review*, *101*(2), 290-300.

[7] Brooks, M. (2014). *At the edge of uncertainty: 11 discoveries taking science by surprise*. Profile Books, 290.

[8] Burgos, J.E. (2018). Selection by reinforcement: A critical reappraisal. *Behavioural Processes*, *161*, 149-160.

[9] Calvin, N.T., & McDowell, J.J. (2015). Unified-theory-of-reinforcement neural networks do not simulate the blocking effect. *Behavioural Processes*, *120*, 54-63.

[10] Chevalley, T., & Schaeken, W. (2016). Considering too few alternatives: The mental model theory of extensional reasoning. *Journal of Experimental Psychology (Hove)*, *65*(4), 728-751.

[11] Clifton, R.K., Freyman, R.L., & Litovsky, R.V. (1994). Listeners expectations about echoes can raise or lower echo threshold. *Acoustical Society of America*, *96*(3), 1533-1535.

[12] Commons, M.L., Navin, J.A., & Davison, M.C. (eds.). (1991). *Signal detection: mechanisms, models and applications*. New Jersey: Hillsdale, 290.

[13] Cooper L.D., Aronson L., Balsam P.D., Gibbon J. (1990). Duration of signal for reinforcement and nonreinforcement in random control procedures. *Journal of Experimental Psychology Animal Behavior Processes*, 16(1), 14-26.

[14] Davison, M., & Jenkins, P.E. (1985). Stimulus discriminability, contingency discriminability, and schedule performance. *Animal Learning & Behaviour*, *13*(1), 77-80.

[15] Domjan, M. (2014). *The principles of learning and behaviour (7th edition).* Wadsworth Publishing, 448.

[16] Dwyer, D.M., Gasalla, P., & Lopez, M. (2019). Partial reinforcement and conditioned taste aversion; No evidence for resistance to extinction. *Quarterly journal of experimental psychology*, *72*(2), 274-284.

[17] Fan, B., Wang, Z., & Wu, F. (2015). *Local image descriptor: Modern approaches*. Springer, 108.

[18] Fulop, S., & Chater, N. (2013). *Why formal learning theory matters for cognitive science*. *Topics in Cognitive Science*, *5*(1), 3-12.

[19] Gershman, S.J. (2015). Do learning rates adapt to the distribution of rewards?. *Psychonomic Bulletin & Review*, 22(5), 1320-1327.

[20] Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Willey, 455.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

[21] Grey, D.A. (1978). Effect of frequency and probability of reward on choice. *Physiological Reports*, 42(2), 543-549.

[22] Hinton, G.E. (2014). Where do features come from?. *Cognitive Science*, *38*(6), 1078-1101.

[23] Honig, W.K., & Staddon, J.E.R. (1977). *Handbook of operant behaviour*. New York: Prantice-Hall, 265.

[24] Kemp, C., Goodman, N.D., & Tenenbaum, J.B. (2011). How to grow a mind: statistics, structure and abstraction. *Science*, *33*(6022), 1279-1285.

[25] Kim, Y.T., Cho, H.C., Seo, J.Y., Leon, H.T., & Klir, G.J. (2010). Intelligent path planning of two cooperated robots based on fuzzy logic. *International Journal of General Systems*, *32*(4), 359-376.

[26] Kinlstom, J.F. (1987). The cognitive unconscious. *Science*, 237(4821), 1445-1452.

[27] Larats, D.B., Biederman, G.B., & Robertson, H.A. (1988). Latent inhibition: attention through response-contingent shock in rats. *The Journal of General Psychology*, *115*(1), 75-82.

[28] LeCun, Y., Bengio, Y., & Hinton, G.E. (2015). Deep learning. *Nature*, 521(7553), 436-444.

[29] Longbing, C., & Philip Y. (Eds.). (2012). *Behaviour Computing: Modelling, analysis, mining, and decision.* Springer, 392.

[30] Lorenz, K. (1981). The Foundations of ethology. *Springer-Verlag Wien*, 380.

[31] Macmillan, N.A., & Creelman, C.D. (2005). *Detection theory: a user's guide*. 2d ed. New York: Lawrance Erlbaum Associates, 520.

[32] McNamara, J.M., & Houston, A.I. (1983). Optimal responding on variable interval schedules. *Behaviour analysis letters, 3*(3), 157-170.

[33] Miltenberger, R.G. (2016). *Behavioural modification: Principles and procedures, 6th Edition.* Thompson Wads worth, 690.

[34] Moreno-Rios, S., Rojas-Barahona, CA., & Garsia-Madruga, J.A. (2014). Perceptual inferences about interminate arrangements of figures. *Acta Psychologica (Amst.)*, *148*(5), 216-225.

[35] Parker, G.B. (2007). Evolving gaits for hexapod robots using cycling genetic algorithms. *International Journal of General Systems*, *34*(3), 301-315.

[36] Pavlov, I.P. (1927). *Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex*. London: Oxford University Press, 142.

[37] Pavlov, I.P. (1932). The reply of a physiologist to psychologists. *Psychological Review*, *39*(2), 91-127.

[38] Ragni, M., & Knauff, M. (2013). A theory and computational model of spatial reasoning with preferred mental model. *Psychological Review, 120*(3), 561-588.

[39] Richards, R.W. (1986). Delay reinforcement and pigeons performance on a one key matching-to-sample task. *Bulletin of the Psychonomic Society*, *24*(1), 85-87.

[40] Saltykov, A.B. (2013). *Functional systems in medicine*. Moscow: Medical informative Agency, 208.

[41] Saltykov, A., & Grachev, S. (2015). *Anticipation and the concept of system-forming factor in the theory of functional systems.* In: Cognitive systems monographs, *25*, 507-520.

[42] Saltykov, A.B., Toloknov, A.V., & Khitrov, N.K. (1989). *Influence of the probability of the random performance of an acquired reaction on the rate of the formation of an instrumental reflex*. Zh Vyssh Nerv Deiat Im I P Pavlova, *39*(4), 654-659.

[43] Saltykov, A.B., Toloknov, A.V., & Khitrov, N.K. (1990). The influence of the probability of random performance of a reaction to the developed on the rate of formation of an instrumental reflex. *Neuroscience and Behavioral Physiology*, *20*(4), 298-303.

[44] Saltykov, A.B., Toloknov, A.V., & Khitrov, N.K. (1992). Frequency of acquired helplessness and initial probability of showing and elaborated reaction. *Patol Fiziol Eksp Ter*, 5-6.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 1786 - 1797
Received 20 January 2021; Accepted 08 February 2021.

[45] Saltykov, A.B., Toloknov, A.V., & Khitrov, N.K. (1993). The optimization of the process of instrumental learning with a low intensity of the conditioned stimulus. *Bulletin of Experimental Biology and Medicine*, *116*(7), 73-75.

[46] Saltykov, A.B., Toloknov, A.V., & Khitrov, N.K. (1998). Optimization of exploratory instrumental activity with low intensity conditioned stimulus. *Bulletin of Experimental Biology and Medicine*, *126*(9), 283-285.

[47] Saltykov, A.B., Smirnov, I.V., Starshov, V.P., & Saltykova, M.M. (1986). Estimation of instrumental conditioning process. *Zh Vyssh Nerv Deiat Im I P Pavlova*, *36*(5), 987-989.

[48] Schindler, C.W., & Weiss S.J. (1982). The influence of positive and negative reinforcement on selective attention in the rat. *Learning and Motivation*, *13*(3), 304-323.

[49] Schonhoff, T.A., & Giordano, A.A. (2006). *Detection and estimation theory and its applications*. New Jersey: Pearson Education, 653.

[50] Scozzafava, R. (2007). Subjective probability versus belief functions in artificial intelligence. *International Journal of General Systems,* 22(2), 197-206.

[51] Simonov P.V. (1991). *Motivated Brain: A neurophysiological analysis of human behavior (Monographs in Physiology).* Rourledge, 280.

[52] Skinner, B.F. (1935). The generic nature of the concepts of stimulus and response. *The Journal of General Psychology*, *12*(1), 40-65.

[53] Skinner, B.F. (1938). *The behaviour of organisms: An experimental analysis*. Oxford, England: Appleton-Century, 457.

[54] Spektor, M.S., & Kellen, D. (2018). The relative merit of empirical priors in non-identifiable and sloppy models: Applications to models of learning and decision-making: Empirical priors. *Psychonomic Bulletin & Review*, *25*(6), 2047-2068.

[55] Staddon, J.E.R. (2016). *Adaptive behaviour and learning: Second Edition*. Cambridge University Press, 618.

[56] Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, *49*(236), 433-460.

[57] Von Helversen, B., Karlsson, L., Mata, R., & Wilke, A. (2013). Why does cue polarity information provide benefits in inference problem? The role of strategy selection and knowledge of cue importance. *Acta Psychologica (Amst.)*, *144*(1), 73-82.

[58] Weiss A., Mirnig N., Bruckenberger U., Strasser E., Tscheligi M., Kuhnelenz B., Wollherr D., Stanczyk B. (2015). The interactive urban robot: user –centred development and final field trial of a direction requesting robot. *Paladyn Journal of Behavioral Robotics*, *6*(1), 42-56.