Logic and Logical Philosophy
Volume 31 (2022), 161–173
DOI: 10.12775/LLP.2021.012

Maciej Malicki

A Formal Analysis of the Concept of Behavioral Individuation of Mental States in the Functionalist Framework

Abstract. The functionalist theory of mind proposes to analyze mental states in terms of internal states of Turing machine, and states of the machine's tape and head. In the paper, I perform a formal analysis of this approach. I define the concepts of behavioral equivalence of Turing machines, and of behavioral individuation of internal states. I prove a theorem saying that for every Turing machine T there exists a Turing machine T' which is behaviorally equivalent to T, and all of whose internal states of T' can be behaviorally individuated. Finally, I discuss some applications of this theorem to computational theories of mind.

Keywords: functionalism; behaviorism; mental states

1. Introduction

By functionalism, I mean a general view, still very prominent in philosophy, psychology, and cognitive science, that mind is, essentially, a computational device, and therefore theories of computation should be used in explaining its nature and workings. One of the first formulations of this stance, commonly called machine state functionalism, was presented by Hilary Putnam in his now classic paper [8]. Putnam proposed identifying the mind with the central unit of a Turing machine, and mental states of the mind with internal states of the machine's central unit in connection with certain states of the machine's tape and head. Thus, for example, pain, understood as a specific mental state related to particular types of stimulus and particular reactions to it, is supposed to be identified with

Received November 18, 2019. *Accepted* July 29, 2021. *Published online* November 8, 2021

some state of the central unit, given that the content of the tape, and the symbols written on it by the head, fit some prescribed pattern.

According to Putnam, the crucial advantage of this model is that it introduces a new level of explanation, which, at least hypothetically, could offer a way out of the vicious circle of reductionist approaches to mental states. This has to do with an old Brentano's thesis which says that intentionality is a fundamentally irreducible phenomenon: every reasonable attempt at explaining (referring to, identifying, etc.) intentional mental states, i.e., states such as believing, expecting, hoping etc., must necessarily involve other mental states. In particular, Roderick M. Chisholm [see 2] and, independently, Peter Geach formulated an argument along these lines in the context of behaviorism, showing that mental states cannot be reduced to behavioral dispositions. Meanwhile, the functionalist model potentially allows for talking about mental states in a consistent, mathematically formalizable manner without reducing them — in terms of inputs, outputs *and other functional states*.

Putnam himself admitted that the idea of using the Turing machine model to explain the nature of mental states was rather vague. He wrote that this "hypothesis schema" should be verified by further theoretical and empirical studies. And in the course of numerous debates on various forms of functionalism that followed, many arguments for and against it were brought forward — for example, the Chinese room argument, the Twin Earth argument (in connection with internalism), etc. In fact, in a later book [9], Hilary Putnam also formulated a critique of his own functionalist program.

However, and rather surprisingly, it seems that so far no one has tried to perform a formal analysis of the hope that modeling the mind as a Turing machine leads to a genuinely non-reductionist perspective. The main goal of this paper is to provide a way of filling up this serious gap. I propose a formalization of the concepts of behavioral equivalence and behavioral individuation of functional states[1]. To be more specific, I define the behavioral disposition of an internal state of a Turing machine as the collection of all possible behavioral configurations, i.e., sequences of

[1] After completing this paper, I found out that a similar approach was proposed in the context of finite automata by E. F. Moore in a seminal paper [5] — see his definition of distinguishable states and Theorem 4. Although Moore's paper was utilized in some discussions related to behaviorism, to the best of my knowledge, no applications of the the concept of distinguishable states have ever been considered in the philosophy of mind.

contents of the tape, and positions of the head that result from a run of the machine starting from this state. This approach seems reasonable — the tape models the environment, i.e., stimulus, and the machine's reactions to it, while the position of the head would represent the localization of the machine in the environment. One could perhaps argue that, in fact, the position of the head should be regarded as part of the machine's internal state of affairs, and its behavior should be restricted only to the content of the tape. This, however, would not affect my conclusions.

In the next step, I define what it means to say that two internal states q and q' of machines T and T' are behaviorally equivalent. This definition captures a rather natural idea that q and q' are identical from the behavioral point of view when they have the same behavioral dispositions, i.e., given any input (and any position of the head), it is not possible to say whether the operating machine is T, starting from the state q, or it is T', starting the from the state q', only by observing the tape and the position of the head. Analogously, the machines T and T' are behaviorally equivalent if their initial states are behaviorally equivalent. And having the notion of behavioral equivalence at hand, it is straightforward to formalize the notion of a behaviorally individuated internal state. Namely, it is a state whose behavioral disposition differs from behavioral dispositions of all other states of the machine, i.e., it can be distinguished from other states just by observing the machine's behavior.

A formalization can be only as good as the insight it gives into the concepts it tries to capture. The one proposed in this paper provides a way for expressing relationships between the mental and the behavioral sides in precise functional terms. For example, Brentano's thesis can be rephrased as follows: internal states are not behaviorally individuated. However, I will try to convince the reader that this framework — especially the notion of a behaviorally individuated state — also supplies effective tools for investigating such relationships. To this end, I will verify the hope that functional descriptions may give a genuinely new level of explanation of mental states. It turns out that a bare functional setup is not enough. This is because — as Theorem 2, stated and proved in the next section shows — for every Turing machine T there exists a Turing machine T' which is behaviorally equivalent to T, and all of whose internal states are behaviorally individuated. Thus, every Turing machine model can be replaced with an equivalent one that is completely describable in a purely behavioral manner.

Obviously, Theorem 2 is a statement about certain formal properties of Turing machines, and not about any particular functionalist theory of mind (be it analytical functionalism saying that mental states, terms, and concepts should be translated into the language of Turing machines, metaphysical functionalism saying that the ontological status of mental states is functional, etc.; see [1] or [7] for a detailed discussion.) Nevertheless, it sheds light on the ingredient of the Turing machine model that seems to play a role whenever — loosely speaking — the internal-external opposition comes up. In order to illustrate it, let me briefly discuss two classical functionalist stances: machine functionalism, and representational theories of mind in the vein of Jerry Fodor.

What are the consequences of Theorem 2 for machine functionalism, i.e., the claim that mental states can be identified with internal states of a Turing machine? Suppose that this hypothesis is correct for some entity E, i.e., there exists a Turing machine T describing the behavior of E, and whose internal states correspond to mental states of E. Now, by the theorem, there also exists a machine T' explaining the behavior of E, and whose internal states are behaviorally individuated. Perhaps it is T, and not T', that is the correct model of E's mind. However, this will not be known until some additional arguments are provided: if there is no independent theory of mind, machines with the same behavioral dispositions cannot be distinguished from one another. In other words, Theorem 2 implies that, in the sole framework of machine functionalism, mental states can be reduced to behavioral states.

Theorem 2 can be also applied to more refined theories of mind. One of the most obvious disadvantages of machine state functionalism is that it does not allow for analyzing and identifying mental states in terms of their content. The so-called representational theories of mind [see Von Eckardt 10] postulate that an (intentional) mental state should not be simply understood as an overall state of the mind but rather as a relation between the mind and a mental representation that forms its (propositional) content: if Jones expects to meet Cecil at the railway station (to consider the example analyzed by Chisholm [2, p. 183]), Jones' mind is in the relation of expecting to the representation 'meet Cecil at the railway station'. In other words, thinking, as well as other mental processes are, in principle, operations performed by the mind on mental representations.

In the computational version of this view, the whole Turing machine, together with its tape, becomes a model of the mind — the central unit models the operations (computations) that can be realized, and the tape is the space, where mental representations are manipulated and stored. Hence, mental processes turn into runs of a Turing machine, and a single mental state is an internal state of the machine, together with an appropriate sequence of symbols that codes the mental representation constituting its content. For instance, Jones expects to meet Cecil at the railway station, when the central unit of Jones' mind is in the internal state 'I expect to . . . ', and the corresponding sequence of symbols on the tape codes the representation 'meet Cecil at the railway station'.

Here, the word "corresponding" is crucial as the tape may contain multiple representations stored by the mind. At this level of generality (which is, obviously, a far-fetched idealization) one can reasonably posit that it is the head of the machine that plays the role of a pointer directing towards appropriate representations. Hence, the position of the head belongs to representation's content.

Now, again as it was the case with machine functionalism, it turns out that it is possible to individuate mental states without referring to internal states of the central unit. Going back to my example, Jones' state of expecting to meet Cecil at the railway station is singled out by the state of the central unit 'I expect to . . . ', and the appropriate part of the tape containing the representation 'meet Cecil at the railway station'. But Theorem 2 says that it cannot be excluded that internal states are identifiable only by their behavioral dispositions (which, this time, should rather be called representational dispositions.) On purely computational grounds, expecting (believing, hoping, etc.) is explicable without necessarily being "led back to the intentional language" [2, p. 185].

Finally, let me comment on the general idea of a realization of a Turing machine that — as one might hope — could help decide which of the machines T and T', given by Theorem 2, is the right candidate for a model of the mind. Even though it is generally recognized that some ingredients of the definition of Turing machine, such as one-dimensional tape, the head that moves only left or right, etc., are not essential to it, still, many scholars maintain that this abstract, mathematical model puts some significant restrictions on its possible (or intended) realizations. Thus, a Turing machine is supposedly 'something like a computer', 'a device performing formal operations on syntactically structured objects', etc. For instance, Jerry Fodor talks about a "classical Turing architecture" [3, p. 31] of the mind, by which he means that the mind is "interestingly like a Turing machine". On the other hand, Steven Pinker [6, p. 3] explains that Turing machines encompass "a variety of systems that we might call 'computational', including ones that perform parallel computation, analogue computation (as in slide rules and adding machines), and fuzzy computation". He is definitely right but still far too modest in his descriptions of what a Turing machine may look like.

In fact, *every* deterministic, and finitary model of mind can be regarded as a realization of a Turing machine, and this is essentially all the insight that can be obtained in this fashion. In particular, simply requiring that internal states of the machine should correspond in some (say, causal) manner to physical states of mind does not even sound like a reasonable starting point for a more detailed analysis. This "technical problem", as Jaegwon Kim [4, p. 88] put it, "something that we will assume can be remedied with a finer-grained notion of an internal state", still forms a fundamental obstacle in developing any mature form of a functionalist theory of mind. Despite many efforts undertaken during the last 50 years, so far there are no convincing candidates for a formal notion of internal structure of the Turing machine.

2. The formalization

According to the definition stated in the Stanford Encyclopedia of Philosophy, a *Turing machine* is a quadruple $T = (Q, \Sigma, q_0, \delta)$, where Q and Σ are finite sets, q_0 is a fixed element of Q, and $\delta : Q \times \Sigma \to \Sigma \times \{L, R\} \times Q$ is a function.

The set Q collects *internal states* of the machine T, i.e., states of the central unit, and q_0 denotes the *initial state*, i.e., the state from which T starts its operation (unless otherwise specified.) The set Σ is the alphabet of possible symbols that can appear on the tape on which T operates, while δ is a *transition function*, which specifies how T operates at every state. I refer to values of δ by

$$\delta(q, s) = (\delta_\Sigma(q, s), \delta_M(x, a), \delta_Q(q, s)).$$

Thus, if the present internal state of the machine is q, and the symbol on the tape at the present position of the head (referred to as the *head value*) is s, then $\delta_\Sigma(q, s)$ is the new symbol written on the tape while the machine proceeds from the state q to the state $\delta_Q(p, a)$, and moves the head to the left, if $\delta_M(q, a) = L$ (here, M stands for 'Move'), or to the right, if $\delta_M(q, a) = R$. As usual, it is also assumed that the tape is blank except for some finite portion of it. This can be formalized by choosing a *blank symbol* $b \in \Sigma$, which is the only symbol that appears infinitely many times on the tape.

For a given Turing machine T, a *configuration* is a finite sequence of the form $\alpha\, q\, \beta$, where α and β are finite words in the alphabet Σ, and $q \in Q$. A configuration $\alpha\, q\, \beta$ encodes a state of the *entire* machine according to the following conventions. First, the non-blank symbols on the tape are $\alpha\beta$. In other words, if $\alpha = \alpha_0 \ldots \alpha_m$, $\beta = \beta_0 \ldots \beta_n$, the tape consists of the following series of symbols: $\ldots bbb\alpha_0 \ldots \alpha_m\beta_0 \ldots \beta_n bbb \ldots$. Second, the central unit is in the state q, and the position of q in the sequence indicates the position of the head on the tape: it reads the symbol β_0. For example, the configuration $q_0 2100$ (here, the alphabet consists of symbols 0, 1 and 2) describes the situation that the central unit is in the initial state q_0, the sequence of non-blank symbols of the tape is 2100, and the head value is 2 — the first non-blank symbol on the tape. Similarly, the configuration $2q_0 100$ indicates that the head value is 1.

Let $q \in Q$ be a fixed internal state. A sequence $c_0, c_1, \ldots$ of configurations is called a *q-trajectory* if q appears in the first configuration c_0, and each configuration c_{n+1} is obtained from the configuration c_n by applying the transition function δ to the internal state and head value coded by c_n. In this way, a q-trajectory codes a complete run of the machine starting from the state q. Note that it is not assumed that q is the initial state. This is because the main rationale behind this notion is to capture also operations of the machine that has already started running, and is presently in some internal state — not necessarily the initial one. The collection of all q-trajectories of T is denoted by the symbol $\mathcal{T}_q^T$, and referred to as the *q-disposition* of T.

Now we introduce the behavioral counterparts of the above notions. For a given configuration $\alpha\, q\, \beta$, the sequence $\alpha * \beta$, where $*$ is just the star symbol replacing the state q, is called a *behavioral configuration.* The idea is to 'forget' about the internal state of the machine (its 'mental state'), and record only the state of the tape (its 'environment'), and the position of the head (the location of the machine in the environment), as indicated by the position of the symbol $*$ in the sequence. For example, the behavioral configuration $2 * 100$ means that the sequence of non-blank symbols on the tape is 2100, and the head value is 1 — however, nothing is known about the internal state of the central unit.

A sequence of behavioral configurations obtained from a q-trajectory is called a *q-behavior*. In other words, a q-behavior is the part of a run of the machine that can be observed 'from the outside': consecutive contents of the tape, and positions of the head — but not its internal states. Also, the collection of all q-behaviors of T is denoted by the symbol $\mathcal{B}^T_q$, and referred to as the *q-behavioral disposition* of T.

Let q be an internal state of a machine T, and let q' be an internal state of a possibly distinct machine T'. The states q and q' are called *behaviorally equivalent* if $\mathcal{B}^T_q = \mathcal{B}^{T'}_{q'}$. This definition formalizes a natural idea that two internal states q and q' are identical from the behavioral point of view if, given the same input at the beginning (and the same position of the head), it is not possible to say whether the operating machine is T, starting from the state q, or it is T', starting the from the state q', only by observing the behavior of the machine. Similarly T and T' are called behaviorally equivalent if the initial states q_0 and q'_0 of T and T', respectively, are behaviorally equivalent. Note that in order to meaningfully define the notion of behaviorally equivalent machines, it is necessary to refer only to their initial states. Otherwise, it could (and would often) happen that a machine was not even behaviorally equivalent to itself.

The above definitions naturally lead to a formalization of the concept of behavioral individuation of internal states. A state q of T is *behaviorally individuated* if no other state of T is behaviorally equivalent to it, i.e., $\mathcal{B}^T_q \neq \mathcal{B}^T_{q'}$ for any $q' \in Q$ with $q \neq q'$. In particular, this definition implies (and, in fact, is equivalent to) the statement that for any other state q', there is an input (and a position of the head) such that the machine's behavior starting from q will, at some point, reveal a difference as compared to its behavior starting from q. Finally, T has *behaviorally individuated (internal) states* if all the states of T are behaviorally individuated.

Now the main technical result of the paper can be stated and proved.

THEOREM A. *For every Turing machine T there exists at least one Turing machine T' which is behaviorally equivalent to T, and has behaviorally individuated internal states.*

Before proceeding to the formal proof, let me comment on the strategy its employs. The main idea is quite simple — the rest is technical machinery required to formally argue that it actually works. I would like to construct a sequence of machines $T_0, T_1, \ldots$, starting with $T_0 = T$, that are behaviorally equivalent to one another, and each machine T_{k+1} is in some sense closer to having behaviorally individuated internal states. I will argue that such a sequence, if appropriately constructed, must terminate, and its last element has behaviorally individuated internal states. This is the machine T' that the theorem postulates.

To be more specific, suppose that I have already constructed such machines $T_0, \dots, T_k$ but T_k does not yet have behaviorally individuated internal states. I select a state q of T_k such that some other state q' is behaviorally equivalent to it. The first case to be considered is that the machine T_k has actually never transitioned to q, i.e., there is no 'link' (specified by the transition function of T_k) from any state r to q. Then, obviously, q can be removed without changing the machine's behavior, so the machine obtained by eliminating q from T_k will be chosen as the new element T_{k+1} of the sequence. Otherwise, there exists a state r such that whenever the machine T_k is in r, and it reads a symbol s from the tape, it is transitioned to q. Now, I can 'rewire' T_k so that, instead of moving from r to q, it moves to q'. As q and q' initially were behaviorally equivalent, it can be shown that such a modification will not alter the machine's behavior. This machine will be the new element T_{k+1}. A crucial feature of the construction is that a single 'link' from the state r to the state q gets removed, a single 'link' from r to q' is added, and, provided that q and q' are selected in a sufficiently careful manner, the sequence must terminate at some point, yielding a machine with behaviorally individuated internal states.

Proof. For a Turing machine $T = (Q, \Sigma, q_0, \delta)$, and a state $q \in Q$, the *indegree* of q is the size of the set

$$\{(r, s) \in Q \times \Sigma : \delta_Q(r, s) = q \text{ for some } s \in \Sigma\}.$$

The indegree of q informs about the number of 'links' from other states to q.

In the first step of the proof, for a given Turing machine $S = (Q, \Sigma, q_0, \delta)$ and internal states $q, q' \in Q$, a new machine S' will be constructed, behaviorally equivalent to S, and obtained from S by removing q, or 'rewiring' it, as informally described above. This construction will be later used to find a sequence $T_0, \dots, T_K$ of behaviorally equivalent machines such that T_K has behaviorally individuated internal states.

Let $S = (Q, \Sigma, q_0, \delta)$ be a Turing machine, and let $q \in Q$ be a state distinct from the initial state q_0. Suppose that $q' \in Q$ is distinct from

q, and is behaviorally equivalent to it. The new Turing machine S' will be defined by specifying a transition function δ', defined over a set of states $Q' \subseteq Q$, and the same alphabet Σ, so that $S' = (Q', \Sigma, q_0, \delta')$ is behaviorally equivalent to S, and either $q \notin Q'$ or the indegree of q calculated in S' is strictly smaller than the indegree of q calculated in S. Moreover, the only state whose indegree can increase in S', as compared to S, is q'.

Suppose that S is never transitioned to q from any state r. In other words, there is no $r \in Q$ and $s \in \Sigma$ such that $\delta_Q(r, s) = q$. Then simply remove q from S, i.e., $Q' = Q \setminus \{q\}$, and define δ' to be δ restricted to the set in $Q' \times \Sigma$.

Otherwise, select some $r \in Q$ and $s \in \Sigma$ such that $\delta_Q(r, s) = q$. Then 'rewire' the machine by setting $\delta''(r, s) = (\delta_\Sigma(r, s), \delta_M(r, s), q')$ (i.e., $\delta''_Q(r, s) = q'$), and $\delta'' : Q \times \Sigma \to \Sigma \times \{L, R\} \times Q$ to be equal to δ for all other arguments. If it so happens that after rewiring, the indegree of q (with respect to δ'') is 0, i.e., the resulting machine is never transitioned to the state q, remove this state, i.e., put $Q' = Q \setminus \{q\}$. Otherwise, put $Q' = Q$. Finally, define δ' to be the restriction of δ'' to $Q' \times \Sigma$.

The machine $S' = (Q', \Sigma, q_0, \delta')$ is as required. Clearly, the indegree of q is strictly smaller in S' than it is in S (because a 'link' from r to q has been removed), and the only state whose indegree increased is q'. Moreover, it is claimed that the machines S and S' are behaviorally equivalent.

In order to prove this claim, it will be shown that if the behaviors of S and S' (starting from a given configuration) are the same as long as the state r has been involved at most n times, they will stay the same as long as r is involved at most $n + 1$ times. From this, it will follows that the behaviors of S and S' are always the same. Select a q-configuration c, and a natural number n.

Let $c_0 = c, c_1, c_2, \ldots$ be a sequence of configurations obtained by applying the machines S' or S to c_k, for $k \geq 0$, in the following way. As long as the state r with the head value s appears in $c_0, \ldots, c_{k+1}$ not more than n times, the machine S' is applied; otherwise S is applied. Let $B_{c,n}$ be the sequence of behavioral configurations obtained from the sequence $c_0, c_1, \ldots$. It is easy to observe that $B_{c,0}$ is just a q-behavior of S: if $n = 0$, the machine S' is never used.

Now it will be shown that $B_{c,n} = B_{c,n+1}$ for every natural n. Select some n. Let $c_0, c_1, \ldots$ be the sequence $B_{c,n}$, and let $d_0, d_1, \ldots$ be the sequence $B_{c,n+1}$. Clearly, if the state r with the head value s is used in constructing $B_{c,n+1}$ at most n times, then $B_{c,n} = B_{c,n+1}$. Otherwise, for some k, the state r with the head value s is used in constructing c_{k+1} — and so d_{k+1} — for the $(n+1)$-th time. Since $c_i = d_i$ for $i \leq k$, the position of the head is the same for c_k and d_k. Then the internal state of S corresponding to c_{k+1} is q, and the internal state of S' corresponding to d_{k+1} is q'. Moreover, by the definition of δ', it holds that $c_{k+1} = d_{k+1}$. As q and q' are behaviorally equivalent, and the machine S is used to construct c_{k+l} as well as d_{k+l} for any $l > 1$, it follows that $c_{k+l} = d_{k+l}$ for any $l > 0$. Thus, $B_{c,n} = B_{c,n+1}$.

In order to finish the proof of the claim, suppose that S and S' are not behaviorally equivalent. Then there exists a q-configuration c such that the q-behavior $c_0, c_1, \ldots$ of S, starting operation from c, differs from the q-behavior $c'_0, c'_1, \ldots$ of S', starting from the same configuration. But this must be witnessed by some behavioral configurations c_k and c'_k (i.e., $c_k \neq c'_k$.) Obviously, the state r was used only finitely many times to construct $c'_0, c'_1, \ldots, c'_k$, so there exists a natural number n such that $B_{c,0} \neq B_{c,n}$. However, in view of the above considerations, this is never the case.

In the second step of the proof, for a given machine T, a machine T' postulated by the theorem will be found, i.e., T' which is behaviorally equivalent to T, and with behaviorally individuated states. Fix a Turing machine $T = (Q, \Sigma, q_0, \delta)$, and fix a linear ordering $\prec$ of internal states of T such that the initial state q_0 is the largest element in this ordering. Put $T_0 = T$, and start building a sequence $T_0, T_1, \ldots$ of Turing machines in the following way. Suppose that $T_0, \ldots, T_k$ have been already constructed. Then select (if possible) distinct behaviorally equivalent internal states q, q' of T_k with the additional requirement that $q \prec q'$ (note that $q \neq q_0$ because q_0 is the largest state.) By applying the above construction to $S = T_k$, q and q', a new machine $T_{k+1} = S'$ is obtained. Continue this procedure as long as possible, i.e., until there are no distinct behaviorally equivalent internal states q, q' to be selected.

Clearly, if the procedure stops, the last machine in the sequence, say T_K, is the machine the theorem postulates: it is behaviorally equivalent to $T = T_0$, and it has no distinct behaviorally equivalent internal states. Therefore in order to finish the proof of the theorem, it needs to be shown that the sequence does terminate at some point.

Suppose it does not, i.e., it is infinite. As there are only finitely many internal states in T_0, by the pigeon-hole principle, there is a state

p and infinitely many numbers k such that p has been selected as q in the construction of T_k. Let p_0 be the smallest such state with respect to the ordering $\prec$. Observe that in this situation, there can be only finitely many k such that p_0 has been chosen as q' in the construction of T_k: otherwise, there would be another state p_1 such that infinitely many times p_1 was chosen as q, and p_0 was chosen as q'. But then $p_1 \prec p_0$, which contradicts the fact that p_0 is the smallest state with respect to $\prec$ chosen infinitely many times as q. However, this implies that, during the construction, the indegree of p_0 was decreased infinitely many times (every time p_0 was chosen as q), while it was increased only finitely many times (every time p_0 was chosen as q'). This is clearly impossible, so the sequence $T_0, T_1, \ldots$ must indeed terminate. □

References

[1] Buechner, J., *Gödel, Putnam, and Functionalism: A New Reading of Representation and Reality*, MIT, 2007. DOI: 10.7551/mitpress/7421.001.0001

[2] Chisholm, R. M., *Perceiving: A Philosophical Study*, Cornell University Press, Ithaca, New York, and Oxford University Press, London, 1957.

[3] Fodor, J. A., *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*, Cambridge, MIT Press, 2000. DOI: 10.7551/mitpress/4627.001.0001

[4] Kim, J. , *Philosophy of Mind*, Westview Press, 1998. DOI: 10.4324/9780429494857

[5] Moore, E. F., "Gedanken experiments on sequential machines", pages 129–156 in C. S. Shannon and J. McCarthy (eds.), *Automata Studies*, Vol. 34, Princeton University Oress, 1956.

[6] Pinker, S., "So how does the mind work?", *Mind and Language* 20, 1 (2005): 1–24. DOI: 10.1111/j.0268-1064.2005.00274.x

[7] Polger, T., *Natural Minds*, A Bradford Book, MIT, 2004. DOI: 10.7551/mitpress/4863.001.0001

[8] Putnam, H., "The nature of mental states", pages 1–223 in W. H. Capitan and D. D. Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh University Press, 1967.

[9] Putnam, H., *Representation and Reality*, A Bradford Book, MIT, 1991. DOI: 10.7551/mitpress/5891.001.0001

[10] Von Eckardt, B., "The representational theory of mind", in K. Frankish and W. Ramsey (eds.), *The Cambridge Handbook of Cognitive Science*, Cambridge University Press, 2012. DOI: 10.1017/CBO9781139033916.004

Maciej Malicki
Institute of Mathematics of the Polish Academy of Sciences
Warsaw, Poland
mamalicki@gmail.com