Philosophical Problems in Science

No 77
2024

Philosophical Problems in Science

Zagadnienia Filozoficzne w Nauce

e-ISSN 2451-0602 (electronic format)
ISSN 0867-8286 (print format, discontinued)

Editorial Office

Philosophical Problems in Science
(Zagadnienia Filozoficzne w Nauce)
e-mail: info@zfn.edu.pl
www.zfn.edu.pl

Publisher: Copernicus Center Foundation
Pl. Szczepański 8, 31-011 Kraków POLAND
tel. (+48) 12 422 44 59
e-mail: info@copernicuscenter.edu.pl
www.copernicuscenter.edu.pl

Philosophical Problems in Science

Zagadnienia Filozoficzne w Nauce

77 — 2024

Articles

Articles

Philosophical Problems in Science (2024), **77**, 3–23
(Zagadnienia Filozoficzne w Nauce)

doi:00000.000

ARTICLE

Explaining the undecidability of first-order logic

Timm Lampert[*†] and Anderson Nakano[‡]

†FernUniversität Hagen
‡Pontifícia Universidade Católica de São Paulo
*Corresponding author. Email: timm.lampert@fernuni-hagen.de

Abstract

Turing proved the unsolvability of the decision problem for first-order logic (*Entscheidungsproblem*) in his famous paper *On Computable Numbers, with an Application to the Entscheidungsproblem*. From this proof it follows that attempts to specify a solution for the *Entscheidungsproblem* through pattern detection in automated theorem proving (ATP) must fail. Turing's proof, however, merely predicts the non-existence of such solutions; it does not construct concrete examples that explain why specific attempts to solve the decision problem by pattern detection fail. ATP-search often runs in infinite loops in the case of unprovable, i.e. not refutable, formulas and one can ask why finite patterns of repeated inference steps cannot serve as criteria for unprovability. We answer this question by constructing pairs of formulas (ϕ, ϕ') such that ϕ is provable (refutable) and ϕ' is unprovable (satisfiable in an infinite domain) but all but the last proof step of the ATP-search for ϕ is a proper part of the endless ATP-search for ϕ'. We generate such pairs of formulas by mimicking computable sequences for a certain kind of universal Turing machine, namely, splitting Turing machines (STMs), via sequences of inference steps in ATP. In contrast to Turing's and the textbooks' method to formalize Turing machines, our method does not rely on further axioms and allows us to transfer the straightforward insight that the halting problem cannot be solved through pattern detection to the case of the *Entscheidungsproblem*. Our method is a constructive alternative to general undecidability proofs that explains why a scientific problem, namely the *Entscheidungsproblem*, is unsolvable in a specific way. This explanation provides a better understanding of the failure of pattern detection, which is of interest to: (i) the programmer, who is concerned with prospects and limits of pattern detection, (ii) the logician, who is interested in identifying logical properties by properties of an ideal notation, and (iii) the philosopher, who is interested in proof methods.

Keywords: constructive proof, automated theorem proving, Entscheidungsproblem, pattern detection, halting problem

1. Introduction

Undecidability proofs of FOL based on Turing machines involve expressing a Turing machine as an FOL formula and demonstrating that the decidability of FOL implies the decidability of some problem that is unsolvable for Turing machines. This latter problem, in turn, is typically proven to be unsolvable using the diagonal method and hypothetical reasoning.

Such a proof is independent of any concrete decision method. It establishes that there cannot exist any algorithm that solves, for instance, the halting problem by referencing the very special case of self-application. This strategy unfolds as follows. Suppose that there exists a machine H that solves the halting problem. Furthermore, consider a machine M that applies H and, after doing so, does not terminate if H decides that it terminates, but terminates otherwise. When M is executed with its own description number as input, this leads to a contradiction, thereby proving the falsity of the initial assumption.

This proof method allows one to prove the non-existence of any algorithm whatsoever that may decide the halting problem and, in consequence, FOL without considering concrete attempts to solve the decision problem. Such a proof method is both powerful and simple, but its explanatory power for the failure of the design of specific procedures is limited (cf. Turing 1937, 246; for reservations against his own method).

In our assessment, part of the dissatisfaction with such proofs stems from the unfulfilled desire to understand more than merely *the fact that* certain decision problems are unsolvable because this would entail contradictory (or tautologous) instructions in the hypothetical diagonal case. In contrast, our aim is *to explain why* a seemingly promising approach to solve the decision problem for FOL, based on pattern detection in ATP, proves futile. Such an explanation must depend on specific features of pattern detection in first-order ATP-search in order to explain what is futile in this approach independent of the general insight that decidability of FOL is impossible due to the diagonal and counterfactual case of self-application. On the one hand, our explanation is more specific and informative than the general insight of undecidability of FOL. On the other hand, it is based on weaker assumptions, as it uses counterexamples to demonstrate the futility of distinguishing refutability and satisfiability by patterns of ATP-search, without relying on counterfactual diagonal cases or expressing Turing machines within FOL.

We believe that explaining undecidability by disproving *specific algorithms* purported to solve decidability is crucial for achieving a deeper understanding of the phenomenon of undecidability. To the best of our knowledge, such explanations are seldom provided. We aspire to alter this situation and thereby contribute to a better understanding of the limitations of ATP and pattern detection in particular and of decidability and algorithmic problem solving in science in general.

We first expound our question in section 2 before presenting our main argument in section 3. Further details of this argument are available in a computer program that we implemented, the behavior of which we briefly describe in section 4.

2. Expounding the question

When designing algorithms for automated reasoning, one of the goals of a software engineer is to develop specific decision procedures that are as powerful as possible. In this context, a logic programmer is not concerned with the diagonalization of a hypothetical, unreal decision procedure. Instead, she may be interested in coming to understand the reasons for limits of actual attempts to spell out concrete decision procedures for FOL formulas. The metalogical literature, however, has predominantly focused on distinguishing decidable from undecidable fragments of FOL, regardless of concrete proof search methods. In doing so, undecidable fragments are typically reduced to expressing problems within FOL that are

undecidable due to hypothetical reasoning and diagonalization (cf. Börger, Grädel, and Gurevich 2001). This approach, however, offers limited insight into the possibilities and constraints of specific methods for making progress in deciding formulas of such fragments. Moreover, for decidable fragments, their decidability may not even be demonstrated by some reasonable decision procedure based on decision criteria (see the following paragraph).

Consider, for example, finite sets of first-order formulas. From a metalogical and classical perspective, we are informed that any finite set of formulas is decidable (cf. Dreben and Goldfarb 1979, p.1). This makes evident the classical, extensional point of view in metalogic. From this point of view, what is asked is whether some decision procedure *exists* rather than *how to specify* a reasonable procedure. As there exists a table with the correct entries 1 and 0 for "provable" and "unprovable" formulas, respectively, for any finite set of formulas, there also exists a computable function that assigns 1 and 0 to each formula. Consequently, in the case of finite sets of formulas, computation is tantamount to "looking up the answer in a table" (Börger, Grädel, and Gurevich 2001, p.239). However, the logic programmer is concerned with developing a program to generate such a table by applying decision criteria. For her, the mere demonstrable existence of such a table, without the means to construct it by applying a decision criterion, is irrelevant.

Therefore, from the perspective of the logic programmer, the relevant question is the extent to which specifying decision criteria is possible for arbitrary, finite or infinite, sets of formulas. From a classical point of view, one may be content with undecidability proofs proving nothing but the absurdity of assuming the existence of a general decision algorithm, independent of considering any decision criteria. However, the more significant question for explaining undecidability concerns the possibility and limits of specific decision criteria. Roughly speaking, the question of undecidability is not an extensional one but rather an intensional one when viewed from this perspective.

Since every provable (refutable) first-order formula can be decided as provable (refutable)—a property known as the semidecidability of FOL—and since every first-order formula with finite models can be decided as satisfiable (and, thus, not refutable[1]), the challenging, albeit still countably infinite, set of formulas consists of those with only infinite models. Formula (1) serves as a simple example of a formula with only infinite models. (2) is its skolemized clause form[2] with the literals numbered[3]. Interpreting Pxy by $x < y$ in the natural numbers yields an infinite model of (1) and (2); see (Börger, Grädel, and Gurevich 2001, p. 33) and (Lampert and Nakano 2020), Theorem 4 for proving that a formula like (1) has only infinite models.[4]

$$\text{FOL formula:} \quad \forall x_1 \exists y_1 (Px_1y_1 \wedge \forall x_2 (Px_2y_1 \vee \neg Px_2x_1)) \wedge \forall x_3 \neg Px_3x_3 \tag{1}$$

$$\text{clause form:} \quad \begin{array}{c} \{\{P[x_1, sk_1(x_1), 1]\}, \{P[x_3, sk_1(x_1), 2], \\ \neg P[x_3, x_2, 3]\}, \{\neg P[x_4, x_4, 4]\}\} \end{array} \tag{2}$$

1. In the following, we basically refer to a proof search for refutablity as is usual in ATP.

2. Skolemization is standard in ATP and allows to eliminate existential quantifiers $\exists\mu$ in favor of skolem-functions, which contain the variables ν such that $\exists\mu$ is in the scope of $\forall\nu$, (cf. Baaz, Egly, and Leitsch 2001, chapter 5.5) and (Nonnengart and Weidenbach 2001, chapter 6.3 and 6.5) for converting FOL-formulas to clauses with so-called outer skolemization that we refer to.

3. Numbering literals serves the purpose of making computation and pattern detection more effective.

4. (Lampert and Nakano 2020) specify a procedure to generate formulas with only infinite models.

Since no finite models are available, model finders provide no assistance for an algorithmic treatment of formulas with only infinite models. For certain fragments of FOL, mere inspection of normal forms suffices to determine the refutability of an initial formula. For example, monadic FOL, which encompasses propositional logic, or so-called Herbrand formulas, which lack disjunction in negated normal form, can be decided without resorting to an exhaustive proof search within a complete calculus. For a straightforward example, consider a disjunctive normal form (DNF) of a propositional formula ϕ: ϕ is refutable if and only if each disjunct of its DNF contains both a literal A and its negation $\neg A$. Similarly, the validity of Aristotelian syllogisms can be read off (and, therefore, decided) from Venn-diagrams.

However, mere inspection of normal forms or, more generally, of FOL-formulas or clauses, is of little help with regard to formulas lacking the finite model property such as the above mentioned decidable FOL-fragments. While finite models can be read off from properties of a finite proof search, infinite models may correspond only to an infinite proof search. Refutable formulas, which have neither finite nor infinite models, exist that are only refutable if some rule is applied iteratively to increase complexity. Examples of such a rule include the so-called rule of expansion in resolution or tableau calculi (cf. the proof search corresponding to Figure 5 below) and $A \vdash A \wedge A$ (or $A \vdash A \vee A$) in other complete calculi of pure FOL. In such cases, one cannot read off the logical property in question, such as refutability, from some normal form expression.[5] Instead, one may need to iteratively increase the complexity of the formula to a certain level to find a proof of refutability. This raises the issue of how to specify the extent of complexity increase. Iterative application of a rule increasing complexity may lead to a proof of a refutable formula or may indicate that no finite model can be inferred from the proof search of a formula with only infinite models.

The most direct and promising method of deciding at least some formulas with infinite models is the so-called "method of saturation". This method consists of a systematic proof search within a complete calculus that yields a proof in the case of provability (refutability) and may terminate in the case of unprovability (satisfiability) due to exhaustive application of the rules of the calculus. The challenge for the logic programmer lies in defining criteria that specify *exhaustive rule applications*, allowing one to conclude that no proof will be found through additional applications. The most direct criterion for this purpose is the so-called *criterion of regularity*. If a sequence of inference steps derives the same formula *twice* on a proof search path, there is no need to continue searching for a proof on this path within an exhaustive search for proofs of minimal length. By this criterion, for instance, the set of formulas that can be converted into prenex normal forms with no existential quantifier in the scope of a universal quantifier can already be decided in tableau or resolution calculi. However, this set of formulas does not include formulas with only infinite models. As soon as existential quantifiers occur in the scope of universal quantifiers—as is the case in formulas with infinite models only—new variables emerge in the iterative application of an inference rule in ATP, rendering regularity insufficient to terminate endless iterations. To address this, a generalization of this criterion is required.

5. Only decidable fragments of FOL can be decided without a rule increasing complexity. (Lampert 2017), e.g., demonstrates how disjunctions of Herbrand formulas can be decided without employing a rule that may increase complexity. In other cases, however, proof search is incomplete without a rule increasing complexity to an arbitrary level.

We distinguish between two senses of *regularity* and, consequently, *regular sequences*: narrow and broad. The former implies the repetition of members in a computable sequence in the strict sense of repeating exactly the same expression (as is the case according to the standard regularity criterion), whereas the latter implies the repetition of a certain *pattern* in a computable sequence that can be identified by a *law*.

By a *law*, we mean a rule that generates, without further computation,[6] potentially infinitely many members of a sequence by generating the *n*th member either directly from previous members (inductive definition) or directly from *n* (explicit definition). Examples of sequences that can be generated by a law include *b*, *ab*, *aab*, *aaab*, *aaaab*, ..., generated by the regular expression "$a*b$"; the sequence of Fibonacci numbers, $0, 1, 0 + 1, 1 + (0 + 1)$, $(0 + 1) + (1 + (0 + 1))$, $(1 + (0 + 1)) + ((0 + 1) + (1 + (0 + 1)))$, . . ., generated by $a_n = a_{n-1} + a_{n-2}$; and the sequence of squares, $1 \cdot 1, (1 + 1) \cdot (1 + 1), (1 + 1 + 1) \cdot (1 + 1 + 1)$, . . ., generated by $a_n = n \cdot n$. An example of a sequence that could not hitherto be generated by a law is the computable sequence of prime numbers. According to our understanding, this sequence, although computable, does not appear to be governed by a pattern that could be identified by a law. Instead of *constructing* the next prime number, we must *search for* it in a finite interval. While we can search for the next prime number within this finite interval, the outcomes of these searches are not governed by a law. As a result, the sequence of prime numbers does not qualify as a regular sequence governed by a pattern in our sense.

Table 1 presents examples in number theory comparing the irregular decimal expansions of $\sqrt{2}$ and $\frac{\pi}{4}$ to their so-called regular ($\sqrt{2}$) and irregular ($\frac{\pi}{4}$) continued fractions, which can be characterized as regular sequences in the narrow and broad senses, respectively. Note that the usual distinction between regular and irregular *continued fractions* does not correspond to our distinction between regular and irregular *sequences*. Instead, it refers to the partial numerators, which are always 1 in the case of regular continued fractions, while they vary in the case of irregular continued fractions. Since the partial numerators are always 1 in regular continued fractions, they are omitted in shorthand notation. Therefore, the shorthand notation for the regular continued fraction for $\sqrt{2}$ is $[1; 2, 2, 2, \ldots]$, which makes evident its regularity in the narrow sense. The irregular continued fraction for $\frac{\pi}{4}$, however, is a regular expansion in the broad sense, as both the numerators and denominators are not strictly identical but develop according to a law.

Henceforth, we will use the unqualified phrase "regular sequence" to refer to a "regular sequence in the broad sense", which is a "sequence generated by a law". Furthermore, whenever we speak of a "pattern of a sequence", we presume that this pattern can be specified by a law. However, this does not imply that this pattern is, in fact, endlessly repeated in a computable sequence; as we will see, this may or may not be the case. That is, we also allow only a part of a finite or infinite sequence to be defined by a law: in this case, *n* in the inductive or explicit definition is, in fact, restricted to a finite number. The pattern itself is always finite but can be repeated either a finite number of times or indefinitely. Finally, when we speak generally of "rules" or "instructions of Turing machines", "Turing machines" or "computable sequences", we do not presume that they are or can be specified by laws. Computable sequences may involve lawless parts or they may involve law-governed finite

6. Note that translations into other notations include further computation. This is why the following sentence in the main text does not present the sequence of Fibonacci numbers or the sequence of squares in the decimal notation, which would translate regular sequences into irregular sequences.

sequences that may be both (i) proper parts of a finite computable sequence or (ii) endlessly repeating parts of an infinite computable sequence. We will argue that this is relevant to our question of the possibility to decide provability based on patterns in ATP.

Table 1. Irregular and regular number sequences

Number	Narrow Regularity	Broad Regularity	Irregularity
$\sqrt{2}$	$1+\cfrac{1}{2+\cfrac{1}{2+\cfrac{1}{\ddots}}}$,	—	1.4142135 . . .
$\frac{\pi}{4}$	—	$1+\cfrac{1^2}{(1+(1+1))+\cfrac{(1+1)^2}{((1+(1+1)+(1+1))+\cfrac{(1+1+1)^2}{\ddots}}}$	0.7853981 . . .

Regularity in the narrow sense is the simplest example of a repeating pattern that enables the termination of a sequence of inference steps in the case of unprovability. If regularity in the narrow sense applies to all open proof paths without a contradictory pair of literals, then the unprovability of the initial input formula can be decided according to the saturation method within an exhaustive proof search. An example of a simple formula that can be decided based on regularity in the narrow sense is (3), with its clause form given in (4), skolem-functions with zero arguments are treated as constants.

$$\text{FOL formula:}\quad \exists y_1 \neg P y_1 y_1 \wedge \exists y_2 \forall x_1 \forall x_2 ((P x_1 x_2 \vee \neg P y_2 x_2) \wedge P y_2 y_2) \tag{3}$$

$$\text{clause form:}\quad \begin{array}{c} \{\{\neg P[sk_1, sk_1, 1]\}, \{P[x_2, x_1, 2], \\ \neg P[sk_2, x_1, 3]\}, \{P[sk_2, sk_2, 4]\}\} \end{array} \tag{4}$$

Formula (3) has a finite model, for instance $\Im(x_1, x_2, y_1, y_2) = \{1, 2\}$, $\Im(P) = \{(1, 1), (2, 1)\}$. Figure 1 illustrates the application of the regularity criterion in the tight connection tableau calculus initialized by clauses with only negative literals. This calculus, along with the ATP-search based on it, is known to be complete (cf. Letz and Stenz 2001). In this paper, we presume this calculus for ATP for clause forms.[7] Additionally, we limit ourselves to cases where the proof tree has only one node (i.e., one tableau), that is, to cases where a *deterministic* proof search is known to be complete.[8] Such a proof search implies that all proofs considered are of minimal length. This is a significant simplification for our reasoning because it enables us to focus on the validity of the purported decision criteria rather than considering the relevance of our criteria for the elimination of nonminimal proofs.

To focus on questions of pattern detection, we employ visualizations of rule application using color diagrams, akin to what is done in the case of, e.g., cellular automata (cf. Wolfram 2002). A configuration of a Turing machine is represented by a sequence of colored squares, where the first square represents the state of the Turing machine and is followed by a sequence

7. Since the focus of our paper is the principal limitation of pattern detection in ATP, we abstain from specifying the technical details of ATP. Interested readers may consult the pertinent paper (Letz and Stenz 2001) and our implementation for details (cf. section 4).

8. Our restriction to translations of deterministic splitting Turing machines into so-called Krom-Horn clauses allows us to do this.

of colored squares representing the symbols on its tape. Different colors represent different states and different symbols. A sequence of configuration is represented by a sequence of sequences of colored squares. Similarly, we represent literals on a proof path with sequences of colored squares: the first square represents the predicate of the literal and is followed by a sequence of colored squares corresponding to the symbols (skolem-functions and variables) at the corresponding argument positions of the predicate. If the last position is a number, this number serves as a counter for initial literals and it is not represented in the color diagram. A branch of the proof consists of a sequence of literals, which is represented as a sequences of sequences of colored squares. For our purposes, it suffices to represent only the main branch of maximal length in our deterministic tableau proofs. Therefore, we can omit the representation of the negation sign, as only negated literals appear on this branch. In Figure 1, e.g., the color diagram represents the main branch $\{\neg P[sk_1, sk_1, 1], \neg P[sk_2, sk_1, 3], \neg P[sk_2, sk_1, 3]\}$: each nth horizontal sequence of colored squares in the diagram corresponds to the n-literal of the main branch, with the predicate represented by the first colored square and the symbols at the argument positions of the predicate represented by the further colored patches. The repetition of the literal $\neg P[sk_2, sk_1, 3]$ in the main branch of Figure 1 is represented by a sequence of colored squares in the second and third horizontal lines of the diagram.

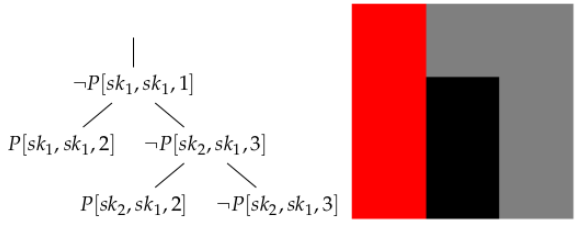

Figure 1. Regularity in the narrow sense in ATP for (4)

Examples suggest that regularity in the narrow sense can be generalized to cases in which the proof search runs in endless loops without repeating expressions in the narrow sense. As Figure 2 illustrates for the simple clause (2) with only infinite models, the proof search may encounter an obvious nesting scenario. We can identify a law that governs how the sequence of formulas develops from the sequence alone, without considering the clauses or instructions of the proof search algorithm. In this case, the sequence of inference steps is, in fact, governed by a repeating pattern, albeit without strict repetition of the same expression. The question is whether we can infer from the finite repetition of a pattern that it will repeat endlessly. We will explain by counter-examples why this question has a negative answer.

There are, of course, more complicated cases of looping than the one shown in Figure 2, and the proof search may become too messy for a human to identify any pattern at all. However, this may be a problem that one may hope to overcome by means of translation into other, more perspicuous notations, or it may be a problem of human pattern identification capabilities that one may hope to overcome by means of machine learning. Thus, one may strive for intricate pattern detection for endless loops in ATP. Our question is how to explain why this endeavor is hopeless. Our philosophical motivation for this question is to explain why a diagrammatic (or "iconic") conception of logic that claims that logical properties, such as provability, are reducible to pattern detection in a suitable notation fails when applied to

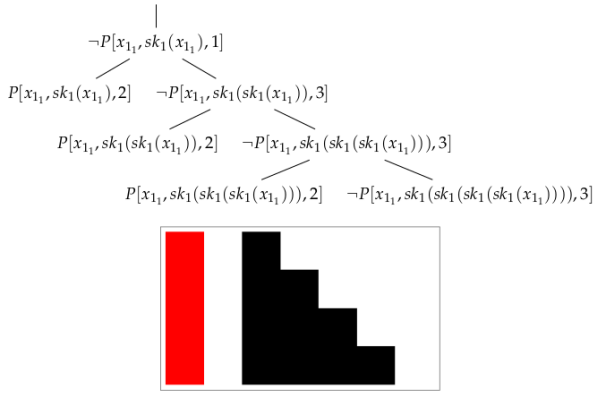

Figure 2. Regularity in the broad sense in ATP for (2)

the whole realm of FOL (cf. Lampert 2018, for details about this diagrammatic (or iconic) conception of logic, which is based on insights from Wittgenstein's early work). We maintain that this conception works only for normal form transformations of *fragments* of FOL (see p. 6 above); we intend to show that and explain by counter-examples why the reducibility claim is incorrect when applied to a general proof search in FOL.

We designate a purported general decision criterion referring to a *finite part* of a computable sequence, wherein this finite part repeats a pattern, as a "loop criterion". This criterion is employed to justify an inference from finite repetitions to endless repetitions and, thus, to determine unprovability. It generalizes the regularity criterion. We aim to answer the following question by providing concrete examples: Why is it impossible to generalize the regularity criterion to correctly determine unprovability by means of a general loop criterion in a complete and automated proof search?

To our knowledge, this question has not been raised in the literature thus far. We shall address this question by considering a simple undecidable class of FOL formulas, namely, the ones which can be expressed as a set of Krom–Horn clauses. Krom–Horn clauses are clauses with at most two members, with at most one being non-negated. Specifically, we will consider only sets of Krom–Horn clauses derived from translations of so-called splitting Turing machines (STMs). We specify STMs for the purpose of making evident the limitations of a decision criterion specified in terms of a loop criterion. By translating STMs into a special sort of Krom–Horn clauses, we can demonstrate how the halting problem for STMs transforms into the decision problem for the corresponding Krom–Horn clauses. Although the deterministic proof search for Krom–Horn clauses will mirror the behavior of STMs, our approach does not involve *expressing* STMs in FOL (as is usually done in undecidability proofs), nor do we employ diagonalization or any axioms (theory) in addition to the pure translation of the input and instructions of STMs. Instead, we mimic the execution of deterministic STMs via a deterministic proof search to show the impossibility of specifying a general criterion for detecting endless loops within an exhaustive proof search in FOL. Our discussion will not only dispense with the diagonal method in combination with hypothetical (counterfactual) reasoning, but will also abstain from *expressing* Turing machines by means of

FOL formulas in relation to their intended interpretation. Rather, it will be purely syntactic and, thus, demonstrable through our implementation of the translation and proof search procedures. In fact, the translations of STMs to Krom-Horn clauses will merely serve as a heuristic to generate Krom-Horn clauses with certain logical properties which can also be proven independent of their relation to STMs.

3. From the halting problem to the *Entscheidungsproblem*

We begin by introducing STMs and explain why an endeavor to solve the halting problem for STMs based on patterns in their evolution is futile (section 3.1). As the main content of our paper, we then show how this explanation transfers to the *Entscheidungsproblem* by considering a proof search for clauses with skolemization (section 3.2).

3.1 Splitting Turing Machines

An STM, or splitting Turing machine, is an automaton equipped with a circular tape consisting of cells that can be split. The specifications and behavior of STMs are defined as follows:

Def. (Splitting Turing Machine, STM): An STM is described by a tuple $S = (Q, \Sigma, f, q_1, q_f, c_n)$, where Q and Σ are the finite sets of states and tape symbols, respectively; $q_1 \in Q$ is the initial state; $q_f \in Q$ is the halting state; c_n is the initial tape of size n that defines the symbols for all n initial positions of the machine; and the transition function f is defined for all $q \in Q$, except q_f, and for all read symbols σ.[9]

We write f as a list of transition rules. Each rule is expressed as a quadruple $t = (q_x, \sigma, \nu, q_y)$, with initial state q_x, read symbol σ, instruction ν, and next state q_y. The possible instructions are as follows:

W*s***:** write the symbol s, $s \in \Sigma$;
S: split the current cell, duplicating its content on the right (clockwise) of the current cell;
L: move the scanner counterclockwise;
R: move the scanner clockwise.

STMs are universal machines because they can simulate clockwise Turing machines (CTMs), which are universal; (cf. Neary and Woods 2009, pp.107-109).[10] Therefore, if the halting problem is solvable for STMs, then it is solvable for all Turing machines.

It is a well-known fact that even Turing machines of minimal complexity can generate rather irregular sequences in their evolution; cf., e.g., Figure 3 for the 5-state, 3-symbol STM described by (5) (the color diagram evolves from left to right, cf. p. 9 for the explanation of color diagrams for Turing machines).

One might wonder whether seemingly irregular sequences already indicate that a decision procedure based on pattern detection for identifying endless looping is futile. However,

9. For simplicity, the definition may omit specifying the value of f for pairs of states and symbols that are never reached.

10. Demonstrating that CTMs can be simulated by STMs is straightforward. Essentially, CTMs are automata that move to the right after every instruction and either write one symbol on the tape or split one cell of the tape and write two symbols in the split cells. These operations are all available in STMs.

Table 2. A 5-state, 3-symbol STM inducing the irregular color diagram in Fig. 3

$$
\begin{array}{ll}
Q\colon & \{P, Q1, Q2, Q3, H\}, q_1 : P, q_f : H, \\
\Sigma\colon & \{0, 1, 2\}, \\
f\colon & \{\{P, 1, \{L\}, P\}, \{P, 0, \{W, 1\}, Q1\}, \{P, 2, \{R\}, Q3\}, \{Q1, 1, \{R\}, Q2\}, \\
 & \{Q1, 0, \{R\}, Q2\}, \{Q2, 1, \{W, 0\}, Q1\}, \{Q2, 2, \{L\}, P\}, \{Q3, 1, \{S\}, Q1\}\}, \\
c_2\colon & \{1, 2\}.
\end{array}
\tag{5}
$$

Figure 3. Color diagram of the evolution of the STM (5) over 150 steps

regularity depends on notation. A different notation may well enhance the possibility of solving decision problems[11], e.g., by converting irregular sequences into regular ones. Approximating a number by means of different sequences in different notations, which can be translated into each other, is one example of this; cf. Table 1. Take the example of square roots: while their representations in decimal notation might not exhibit a discernible pattern, their regular continued fractions demonstrate *periodicity*, allowing for easier identification. Similarly, one could argue that by translating the irregular sequence generated by an STM into a regular one through notation transformation, a recognizable pattern might emerge. Our approach of translating STMs into clauses or FOL formulas, and subsequently generating sequences of inference steps by applying a logical calculus, allows us to explore the relations between sequences in different notations with respect to the resulting patterns. The normal form transformations employed to solve decision problems for fragments of FOL, as discussed earlier (cf. p. 6), illustrate how the ability to solve decision problems in logic depends on a notation that may reveal logical properties in the form of common patterns. From the practical standpoint of a software engineer, one might also wonder whether machine learning could outperform humans and gain the ability to decide problems (at least to a reliable degree) by learning from patterns in a training database.

However, it can be demonstrated through a representative example that attempting to solve the halting problem based on patterns of STM sequences or to solve the *Entscheidungsproblem* based on patterns of sequences of inference steps is futile, irrespective of the extent to which irregular sequences can be transformed into regular ones. We can demonstrate this by considering *regular* instead of irregular sequences, i.e., by considering cases in which a pattern is *indeed* found. We first do this for sequences computed by STMs and the question of solving the halting problem and then ask whether the same approach can be extended to sequences in ATP and the *Entscheidungsproblem*. Thus, we do not explain undecidability due to the lack of detectable patterns in unsuitable notations. Instead, we explain undecidability due to the impossibility of distinguishing properties such as halting / non-halting in the case of STMs and refutability / satisfiability in the case of FOL *despite of the identification of patterns*.

11. Cf. (Lampert 2020) for a general discussion of this claim.

Our example, in which STM1 and STM2 are specified as shown in Table 3 and their color evolution diagrams are presented in Figure 4, demonstrates that mere repetition of a regular pattern is not sufficient to decide that a Turing machine is nonhalting. STM2 differs from STM1 solely by the replacement of one of the halting instructions with a nonhalting instruction. Their evolution diagrams remains identical for the initial 78 steps, encompassing a complete repetition of a regular pattern. However, STM1 halts in the next step, while STM2 continues forever in a regular manner.

Table 3. STM1 (halting) and STM2 (not halting), differing by one instruction

Q:	$\{P, P1, P2, P3, P4, H\}, q_i : P, q_f : H,$
Σ:	$\{1, 2, 3, 4\},$
f:	$\{\{P, 1, \{W, 2\}, P1\}, \{P, 2, \{R\}, P\}, \{P, 3, \{R\}, P4\},$
	STM1:$\{\mathbf{P}, \mathbf{4}, \{\mathbf{W}, \mathbf{4}\}, \mathbf{H}\}$, **STM2:**$\{\mathbf{P}, \mathbf{4}, \{\mathbf{W}, \mathbf{4}\}, \mathbf{P1}\},$
	$\{P1, 1, \{R\}, P1\}, \{P1, 2, \{R\}, P1\}, \{P2, 3, \{R\}, P\}, \{P1, 4, \{R\}, P\},$
	$\{P2, 1, \{R\}, P2\}, \{P2, 2, \{W, 1\}, H\}, \{P2, 3, \{R\}, H\}, \{P2, 4, \{R\}, P3\},$
	$\{P3, 1, \{S\}, P\}, \{P3, 2, \{S\}, H\}, \{P3, 3, \{R\}, H\}, \{P3, 4, \{W, 1\}, H\},$
	$\{P4, 1, \{R\}, P4\}, \{P4, 2, \{W, 1\}, P4\}, \{P4, 3\{R\}, P2\}, \{P4, 4, \{R\}, P4\}\},$
c_5:	$\{1, 3, 1, 1, 4\}.$

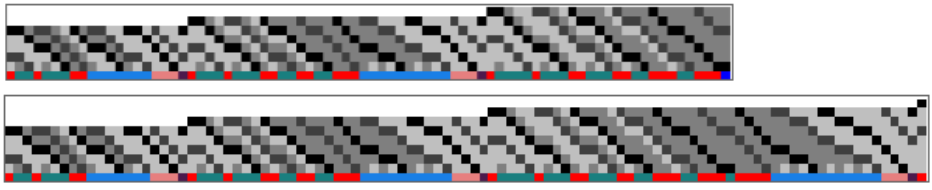

Figure 4. Color diagrams for STM1 and STM2 up to 100 steps

It is easy to describe the difference of the functions that the instructions of STM1 and STM2 implement. Both take as input two sequences of 1s, a and b, separated by the symbols 3 and 4. While a increments by 1 in each iteration, b remains constant. STM1 halts when $a > b$, whereas STM2 goes on with incrementing a forever. Yet, describing the implemented functions does not amount to provide a general decision criterion for distinguishing finite from infinite loop processes. Recall that, from the point of view of metalogic, it is possible to mechanically decide for any members of an arbitrary *finite* set of STMs (that may well include STM1 and STM2) whether they halt or do not halt, cf. p. 5. It is even possible to devise an algorithm that decides, for an infinite set of STMs that includes STM1 and STM2, whether its members halt or do not halt. This can be done by considering special forms of machines, e.g., machines that implement Do-until loops that do not modify the loop counter during the loop, but only increment or decrement the counter after each loop until a certain value is reached and, additionally, each single loop is known to end after a finite number of steps. Yet,

what is in question is a decision criterion that applies to *all* STMs, thereby distinguishing finite from infinite loop processes *in general*, solely based on detecting repeating patterns, regardless of an upper bound on loop iterations. This becomes imperative when dealing with STMs featuring While-loops where termination hinges not on reaching a predefined counter but on satisfying a specific condition. Our pair of machines, STM1 and STM2, epitomizes the fundamental challenge of formulating a *general* criterion for escaping a loop by pattern detection. Unlike the counterfactual diagonal case, such a pair comprises concrete STMs, which enables the translation of the problem to a factual pair of expressions within FOL and its illustration by real automated proof search.

In investigating sequences, we assume that it is possible to identify mechanically loops that repeat a certain pattern. In the case of STM1 and STM2, e.g., the looping process starts from configurations differing solely in the increasing number of 1s in the first sequence a of 1s. Within each iteration of the loop, the same instructions are applied in the same order, with only the number of applications of these instructions varying in each repetition due to the increasing nesting. A universal machine can identify the looping process by detecting the successive increase of nesting and the repeating succession of rule applications resulting from it. Given these assumptions, the question arises whether it is possible to infer non-halting behavior from the identification of a looping process.

Pairs of STMs such as STM1 and STM2 show that and explain why the answer to this question is negative. These machines share a looping process; yet, while STM1 escapes the loop and halts, STM2 repeats the looping process endlessly. All of the halting instructions in both STM1 and STM2, except the one absent in STM2, are irrelevant as their conditions will never be met. STM2 does the same as STM1 with the only difference being that it continues its execution after each comparison of the increasing number a with b. A loop criterion is refuted by the fact that, by simply redefining a relevant instruction where a halting condition will be satisfied, one can specify a machine that behaves likewise but continues its execution while the original halts. Clearly, examples such as STM1 and STM2 can be extended to an arbitrary number of pairs of halting and nonhalting machines by redefining relevant halting instructions.

Our refutation implies that STM2 can no longer be decided as nonhalting according to a pure loop criterion that is based on nothing but the detection of a pattern within a *finite* part of a computable sequence. Note that, by increasing the second number b against which the first number a is compared, we can arbitrarily increase the number of repetitions executed before STM1 comes to a halt. Therefore, a loop criterion cannot be rescued by *merely* increasing the number of repetitions under consideration.

We do not refute a loop criterion by objecting that it may be challenging to identify patterns in a suitable notation. Instead, our refutation stems from the fact that such a criterion fails to provide a sufficient condition for nonhalting when loops can be mechanically identified. In what follows, our aim is to extend this straightforward refutation of the validity of a loop criterion for solving the halting problem to the attempt of applying such a criterion to solve the *Entscheidungsproblem* within ATP. In this context, it is not equally straightforward to see the failure of a loop criterion when employed as a general decision criterion. Since regularity and decidability depend on notation, refuting a loop criterion for sequences generated by STMs does not *immediately* imply refuting it for ATP. To establish this, we must demonstrate how the (trivial) impossibility of solving the halting problem based on a loop criterion can

be transposed to show the (nontrivial) impossibility of solving the *Entscheidungsproblem* based on a loop criterion.

Furthermore, proofs are typically examined in logic without considering their relation to the execution of Turing machines. One may even harbor reservations concerning the endeavor to overload FOL formulas with interpretations that go beyond a pure proof-theoretic point of view. Consequently, one may be inclined to abstain from *expressing* Turing machines by means of FOL formulas by making use of interpretations $\Im$ of FOL-expressions intending to capture the states, the tape and the instructions of TMs.[12] To circumvent these reservations, we will show in the following that and how the impossibility of specifying a loop criterion for the halting problem can be extended to the impossibility of doing this within ATP without *expressing* STMs within the language of FOL. Instead, we will show how to mimic the behavior of STMs in a deterministic proof search for the rather simple case of Krom–Horn clauses.

3.2 ATP-search

We proceed to show how our explanation of the invalidity of a loop criterion as a decision criterion for the halting problem can be extended to establish the unsolvability of the decision problem for FOL by employing a loop criterion. For this sake, we specify an ATP-search that emulates the behaviour of STMs. This can be accomplished by utilizing the well-known ATP-search within a tight connection tableau calculus, which is identical to an ATP-search in the also well-known linear resolution calculus for the special case of Krom-Horn clauses that we consider, cf. footnote 13. The crucial point is that the ATP-search, in fact, produces repeating patterns in terms of regularity not only in the narrow but also in the broad sense. Thus, we demonstrate that increasing complexity through the endless iterative application of inference rules, which is a necessary feature of any complete calculus for FOL, cannot be circumvented by a loop criterion, even in the case where the ATP-search generates discernible repeating patterns that would enable the application of such a criterion. Note that mimicking the behaviour of TMs by ATP-search is not a trivial task: Our translation of STMs into Krom-Horn clauses is tailored to specifically emulate the behaviour of a certain type of TMs by a particular ATP-search. Different types of TMs (e.g., those without splitting cells), other notations of FOL-expressions (e.g., without skolemization) and other correct and complete calculi (e.g., those requiring iterative application of $\wedge$I) may well render the application of a loop criterion impractical, if not impossible. In these cases, decidability through pattern detection is thwarted due to the absence of discernible or repeated patterns. However, this is a feature of notation that one can hope to overcome by the reduction to proper notations and calculi. We show that even in the case of such a reduction, a loop criterion remains unreliable.

Let us first enumerate the elements we will use for representing the machine and the contents of its tape:

1. A Skolem constant sk_G.
2. A Skolem constant sk_s for every $s \in Q$.
3. A unary Skolem function sk_a for every $a \in \Sigma$.

12. Cf. (Lampert 2020), section 5, which criticizes semantic versions of undecidability proofs of FOL. Our purely syntactic version of mimicking the behavior of STMs circumvents these concerns entirely.

4. A predicate letter P_s with arity $n+1$ for an STM with n initials cells for every $s \in Q$.
5. Some auxiliary predicate letters, to be defined below.

An STM in a certain state and with certain symbols written on its tape will be represented by a literal (i.e. an atomic formula or its negation) as follows: the different states of the machine will be represented by different predicate letters (and, additionally, by different Skolem constants, see 2 above; this redundancy is introduced merely for simplicity); the tape of the machine will be represented by the arguments of the predicates, with the first argument representing the position of the scanner of the machine; the different symbols that may be written on the tape of the machine will be represented by different Skolem functions; finally, the potentially infinite nature of the tape (due to the splitting operation) will be represented by the potentially infinite nesting of these Skolem functions. The Skolem constant sk_G marks the end (the "ground") of nesting.

From here on, we will employ the metavariable α to represent variables that run through the values of Σ and the metavariable β to represent variables that run through the values of Q. Additionally, we stipulate that the first argument of the predicates will always represent only a *single* cell in the machine. The reason for this stipulation is merely practical.

We assume that the tape is initially filled with $s_1, s_2, \cdots, s_n$, where all $s_i \in \Sigma$. Given that we utilize the tight connection calculus beginning with a single negated literal, cf. p. 8, our initial clause contains only one negated literal and our clauses of length 2 will start with a positive and end with a negated literal. The initial clauses of the FOL formula in clause normal form are as follows:

$$\{\neg F_{aux0}[sk_G, ..., sk_G]\} \tag{6}$$

$$\{F_{aux0}[x_1, \ldots, x_1], \neg P_{q_1}[sk_{s_1}(x_1), sk_{s_2}(x_1), \ldots, sk_{s_n}(x_1), sk_G]\}. \tag{7}$$

The arity of F_{aux0} is n, i.e. the number of initial cells. Note that the tape, which is of size n, is represented by $n+1$ arguments in predicate letters P_s such as P_{q_1}. The $(n+1)$-th argument will always be sk_G, which is useful for the specification of Rule R below.

The ATP-search starts by deriving $\neg P_{q_1}[sk_{s_1}(sk_G), sk_{s_2}(sk_G), \ldots, sk_{s_n}(sk_G), sk_G]$ from (6) and (7). We could have simplified (6) and (7) to this literal. However, we want our clauses to be easily translatable into pure FOL-expression without skolemization, cf. section 4. For this sake, skolem functions should not occur within skolem-functions prior to substitutions in the initial clauses, i.e. prior to substitution of variables during the application of inference rules. For the same reason, we impose the general restriction that, in the clauses, the only permissible argument for the Skolem functions sk_a for every $a \in \Sigma$ is x_1.

For the halting state q_f, we add the following unit clause:

$$\{P_{q_f}[x_1, ..., x_n, sk_G]\}. \tag{8}$$

Let us now see which clauses are needed for every kind of transition rule. We assume that the description of each rule is given in the format $(q_x, \sigma_1, \nu, q_y)$.

We will omit the translation of Rule L since it is analogous to Rule R and since we do not need it in our examples or for our argument. Rules W and S are straightforward and obviously yield a literal representing the resulting state and tape of the corresponding STM.

3.2.1 Rule W

Suppose that the symbol to be written is σ_2. To represent Rule W, we simply add the following clause:

$$\{P_{q_x}[sk_{\sigma_1}(x_1), x_2, x_3, \ldots, x_{n+1}], \neg P_{q_y}[sk_{\sigma_2}(x_1), x_2, x_3, \ldots, x_{n+1}]\}. \tag{9}$$

3.2.2 Rule S

To represent Rule S, we add the following clause:

$$\{P_{q_x}[sk_{\sigma_1}(x_1), x_2, x_3, \ldots, x_{n+1}], \neg F_{auxS_{\sigma_1}}[sk_{\sigma_1}(x_1), x_2, x_3, \ldots, x_{n+1}, sk_{q_y}]\}. \tag{10}$$

Additionally, we include the following clauses in the set of clauses (these clauses are included only once, not for every Rule S dictating the machine's behavior):

$$\{F_{auxS_\alpha}[x_2, x_1, x_3, \ldots, x_{n+1}, sk_\beta], \neg P_\beta[x_2, sk_\alpha(x_1), x_3, \ldots, x_{n+1}]\}. \tag{11}$$

The rationale for the introduction of the auxiliary predicate letters F_{auxS_α} is that they allow for the representation of the splitting of the cells while obeying the restriction that, in the clauses, the only permissible argument for the Skolem functions sk_a for every $a \in \Sigma$ is x_1, cf. p. 16.

3.2.3 Rule R

The specification of Rule R is considerably more intricate compared to Rules W and S. Let us begin with an example to illustrate how mimicking this rule in a tight connection tableau calculus unfolds. Imagine that at a given instant during the machine's execution, its state q_x and the tape $[1, 2, 3, 4, 5, 6]$ are represented by the following literal, which appears on a leaf of a certain open branch of the tableau:

$$P_{q_x}[sk_1(sk_G), sk_2(sk_3(sk_4(sk_G))), sk_5(sk_6(sk_G)), sk_G].$$

The idea is that at the end of the mimicking operations in the tableau, we shall obtain the following literal on the leaf of this branch:

$$P_{q_y}[sk_2(sk_G), sk_3(sk_4(sk_5(sk_6(sk_G)))), sk_1(sk_G), sk_G].$$

representing the tape $[2, 3, 4, 5, 6, 1]$ and the machine in state q_y. This evolution during the application of the rules of the tableau calculus is designed to mimic the fact that the header of the machine moves one position to the right on the tape, transitioning from the state q_x to the state q_y.

To ensure that Rule R will be applied only if the scanned symbol is σ_1, we include the following clause:

$$\{P_{q_x}[sk_{\sigma_1}(x_1), x_2, x_3, ..., x_{n+1}\}, \neg F_{aux_R}[sk_{\sigma_1}(x_1), x_2, x_3, ..., x_{n+1}, sk_{q_y}]\}. \tag{12}$$

Now, we start moving the symbols to the right. Initially, we add the following clauses to the set of clauses:

$$\begin{gathered}\{F_{aux_R}[x_2, sk_\alpha(x_1), x_3, \ldots, x_{n+2}], \\ \neg F_{aux1_\alpha}[sk_\alpha(x_1), x_1, sk_G, x_3, \ldots, x_n, x_2, x_{n+1}, x_{n+2}]\}.\end{gathered} \tag{13}$$

The rationale for the arguments of F_{aux1_α} will only be made clear after the following operations 1 to 4 are explained. We need to perform the following operations with the arguments of F_{aux1_α} to represent the tape after the execution of the rule:

1. Set the first argument to $sk_\alpha(sk_G)$ (for the particular value of α in question).
2. Pop the represented symbols on the tape contained in x_1 (second position of F_{aux1_α}), push them over the third position, and finally remove the second position.
3. Pop the inverted x_1 obtained in operation 2 above and push these symbols over x_3 to put them in the right order, again removing the second position.
4. Use the last argument of F_{aux1_α}, which stores the next state, to finally construct the correct predicate letter.

Regarding clauses (13)–(21), and unlike clause (12) above, we anticipate that these clauses do not need to be added for *every* Rule R. Instead, we include them only once.

To accomplish the first task, we include the following clauses:

$$\{F_{aux1_\alpha}[x_2, x_3, \ldots, x_{n+3}, x_1, x_{n+4}], \neg F_{aux2}[sk_\alpha(x_1), x_3, \ldots, x_{n+3}, x_1, x_{n+4}]\}. \tag{14}$$

To accomplish the second task, we include the following clauses:

$$\{F_{aux2}[x_2, sk_\alpha(x_1), x_3, \ldots, x_{n+4}], \neg F_{aux3_\alpha}[x_2, x_1, x_3, \ldots, x_{n+4}]\} \tag{15}$$

$$\{F_{aux3_\alpha}[x_2, x_3, x_1, x_4, \ldots, x_{n+4}], \neg F_{aux2}[x_2, x_3, sk_\alpha(x_1), x_4, \ldots, x_{n+4}]\} \tag{16}$$

$$\{F_{aux2}[x_1, sk_G, x_2 \ldots, x_{n+3}], \neg F_{aux4}[x_1, x_2 \ldots, x_{n+3}]]\}. \tag{17}$$

The third task is similarly completed by including the following clauses:

$$\{F_{aux4}[x_2, sk_\alpha(x_1), x_3, \ldots, x_{n+3}], \neg F_{aux5_\alpha}[x_2, x_1, x_3, \ldots, x_{n+3}]\} \tag{18}$$

$$\{F_{aux5_\alpha}[x_2, x_3, x_1, \ldots, x_{n+3}], \neg F_{aux4}[x_2, x_3, sk_\alpha(x_1), \ldots, x_{n+3}]\} \tag{19}$$

$$\{F_{aux4}[x_1, sk_G, x_2 \ldots, x_{n+2}], \neg F_{aux6}[x_1, x_2 \ldots, x_{n+2}]\}. \tag{20}$$

Finally, to accomplish the fourth task, we include the following clauses:

$$\{F_{aux6}[x_1, \ldots, x_{n+1}, sk_\beta], \neg P_\beta[x_1, \ldots, x_{n+1}]\}. \tag{21}$$

3.2.4 Mimicking the evolution of STMs

Applying the rules of our translation procedure to a given STM yields a set of Krom–Horn clauses. The initial and final states are translated into clauses of length 1, while all instructions

are translated into clauses of length 2 with exactly one non-negated literal. Our presumed complete ATP procedure based on tight connection tableau, initiated with clauses containing only negated literals, makes it possible to mimic the execution of deterministic STMs via a very simple deterministic ATP process. It starts with the initial clause, since this is the only clause that contains only negated literals. Subsequently, there is precisely one expansion step to be iteratively executed (this fact readily follows from inspection of the provided set of clauses). Reduction steps, which do occur in a complete tableau proof search for the whole realm of FOL, are absent in the complete tableau proof search for the special case of Krom–Horn clauses. The proof search terminates if and only if the halting state is reached. Since the proof procedure is deterministic, any proof is necessarily of minimal length. Incidentally, the same is true for a proof search within the linear resolution calculus, which does not differ significantly from the proof search in tableaux in the special case of Krom–Horn clauses.[13]

Although ATP mimics the evolution of STMs, there is no one-to-one correspondence but rather a one-to-many correspondence between the steps of the STMs and the steps in our tableau calculus. The primary reason for this is Rule R, which entails a pop–push process for the Skolem functions in positions two and three of the predicates to achieve the correct order of the arguments in the Skolem functions. However, one can compare the literals in the ATP process following each translation of an instruction with the tape after the performance of an instruction during STM execution. This can be seen, for example, by contrasting the color diagrams of the STM sequences with those of the ATP process; see, e.g., Figure 4 and Figure 5 (cf. p. 9 for the explanation of color diagrams).

The translation T(STM1) of STM1, which is a provable formula, differs by only one literal in one clause from the translation T(STM2) of STM2, which is a satisfiable formula. The automated proof of T(STM1) in tableaux takes 634 steps. Figure 5 shows the evolution diagrams for T(STM1) and T(STM2), divided into three parts (the partial diagrams are meant to be read from top to bottom). The outermost left partial diagram covers steps 1–96, while the subsequent partial diagram encompasses steps 97–346 after the first splitting. The third partial diagram shows steps 347–634 for T(STM1) after the second splitting, while the last partial diagram depicts steps 347–640 for T(STM2) after the second splitting. The third and fourth partial diagrams diverge only at step 634, which is the last step in the proof of T(STM1), while the proof search for T(STM2) continues indefinitely. The patterns repeat after each splitting, becoming more nested; this extends the pop–push processes without causing new symbols to be written or new states to be entered relative to the previous sequence of inference steps within one splitting period. Thus, similar to the computable sequences of STM1 and STM2, the sequences of inference steps for T(STM1) and T(STM2) share the same repeating pattern. However, this pattern endlessly repeats only in the proof search for T(STM2). This demonstrates that the invalidity of the loop criterion for deciding whether halting occurs in the case of STM1 and STM2 extends to the invalidity of deciding provability for T(STM1) and T(STM2) based on a loop criterion for an automated proof search within a tableau or resolution calculus.[14]

13. Full resolution reduces to binary resolution, factorization is not needed, the expansion rule and binary resolution are identical in this case, and (Henschen and Wos 1974) have shown that linear *input* resolution is complete for Horn clauses.

14. T(STM1) and T(STM2) each contain 79 clauses and thus are too long to be printed here. All automatically generated diagrams and translations from STM1 and STM2 can be viewed at github.com/TimmLampert/KromHornSolver. The printed output of the color diagrams and their symbolic expressions is 125 and 133 pages long.

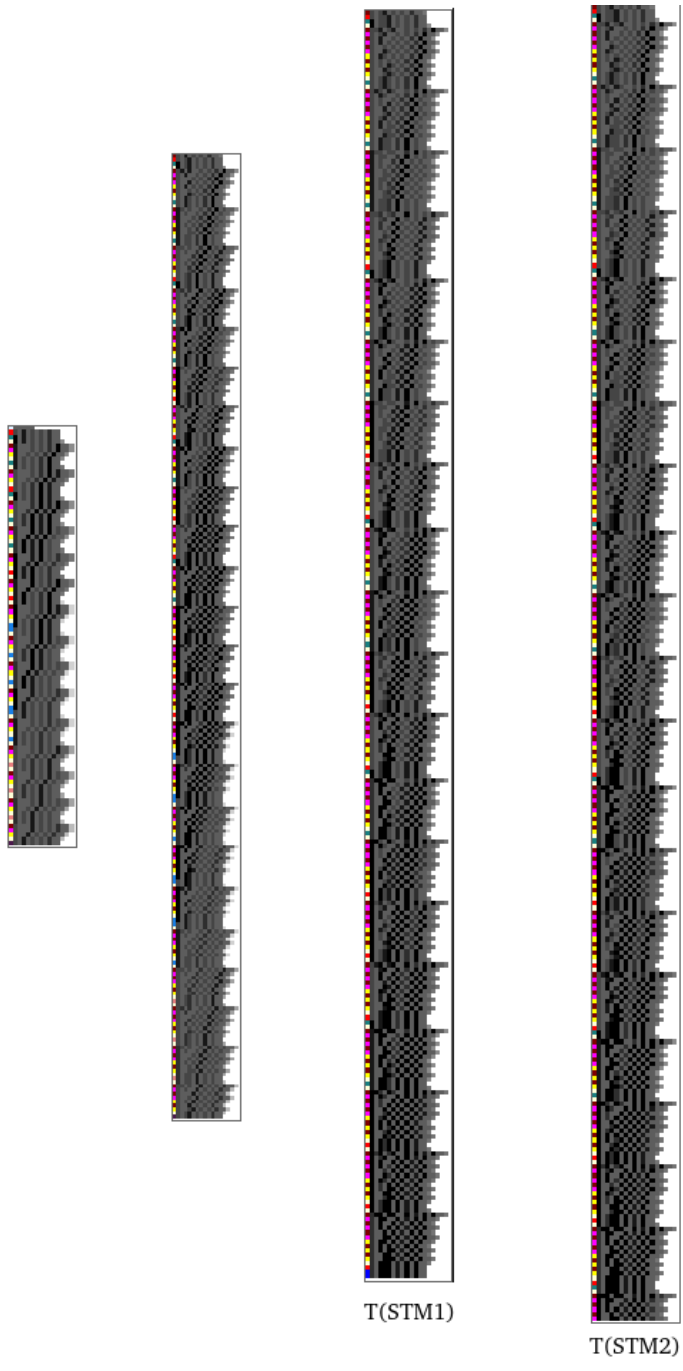

Figure 5. Color diagrams of the proof searches for T(STM1) and T(STM2), showing that although both are governed by the same pattern before the proof search for T(STM1) terminates, only the proof search for T(STM2) continues to repeat this pattern.

Since we can produce an arbitrary number of further pairs of STMs for this case, we can also generate an arbitrary number of further pairs of provable and satisfiable formulas that share a regular sequence of inference steps up to an arbitrary length. This demonstrates the futility of specifying a loop criterion for ATP-search.

4. KromHornSolver

In contrast to metamathematical proof methods, such as the diagonal method, we base our results on examples and computation. STMs and their automated translations into Krom–Horn clauses are especially suited for explaining undecidability in terms of pattern detection in ATP. To aid the visualization of our argument, we compute color diagrams instead of printing complex symbolic expressions and proofs.

Our results are based on a *Mathematica* program called *KromHornSolver*, which we implemented ourselves.[15] This program not only conducts proof searches for clauses but also for corresponding FOL-formulas without skolemization, which are generated from the clauses. The ATP-proof search for those FOL-formulas is based on a different calculus that applies the rule $\wedge$I instead of the expansion rule prior to universal quantifier eliminations. Avoiding skolemization makes it impossible to reproduce the repeating patterns of the evolution of STMs in the case of splitting cells, thereby precluding the application of a loop criterion. This occurs due to the fact that the splitting tape can no longer be represented by P-literals with increasing nested skolem-functions but only by several P-literals dispersed among other literals in different scopes of different quantifiers, which increase in number by the splitting process. Consequently, there is no remaining main branch of an ATP-search with P-literals corresponding to the evolution of the STM. Instead, the evolution of STMs is encoded in a complex FOL-expression with literals not arranged in a regular sequence but distributed across different scopes of different quantifiers. By implementing two different ATP-searchs based on two different notation (with and without skolemization), we intended to underscore the impact of notation and calculus on pattern detection.

The *KromHornSolver* takes an STM plus an upper bound for execution as its input. It then performs the following steps:

Step 1: The array for the steps of the STM is generated its color diagram is printed.
Step 2: The STM is translated into clauses with Skolem functions and into a pure FOL formula (without Skolem functions).
Step 3: The deterministic tableau is generated for the clauses and its color diagram is printed.
Step 4: The tableau proof is translated into a pure FOL formula encoding the corresponding proof in proof search without skolemization and its color diagram is printed.
Step 5: Attempts to specify a loop criterion are checked.

Step 5 refuted the correctness of a loop criterion for the ATP-search for clauses with skolemization, while the corresponding ATP-search for pure FOL-formulas without skolemization does not allow for specifying a general loop criterion due to the lack of regular sequences. In both cases, it becomes impossible to decide satisfiability by identifying repeating patterns.

15. This program is available at www.github.com/TimmLampert/KromHornSolver, and an implementation of it can be accessed at www2.cms.hu-berlin.de/newlogic/webMathematica/Logic/kromhornsolver.jsp.

5. Conclusion

Our main concern is to *demonstrate that* and to *explain why the particular putative attempt* to solve the *Entscheidungsproblem* within *given* calculi and proof search algorithms by adding a general loop criterion is futile, rather than *proving in general that* the *Entscheidungsproblem* is unsolvable *by any method*. Our explanation is based on mimicking STM sequences in ATP. However, this does not mean that our explanation hinges on the translation of STMs into FOL. We merely use this translation method as a heuristic to generate pairs of sequences in ATP that share a repeating pattern. Once these sequences are generated, the validity of a loop criterion for an ATP-search yielding repeating patterns is directly refuted by the fact that an infinite sequence for an unprovable formula and a finite sequence for a provable formula contain the same repeating pattern.

Our discussion of the limits of a loop criterion can be seen as an instance of the general insight that finite, regular parts of computable sequences lack an unambiguous continuation. However, we argue that simply referring to this insight in a general context is insufficient when discussing ATP. Instead, one must prove that and how such a general statement does apply to the problem of specifying decision criteria within ATP via pattern detection. We achieve this by employing the *KromHornSolver* and applying it to cases such as the translations of STM1 and STM2 in our example. This approach allows us to illustrate and substantiate the implications of the broader insight within the context of ATP.

References

Baaz, Matthias, Uwe Egly, and Alexander Leitsch. 2001. Normal Form Transformations. In *Handbook of Automated Reasoning Vol. I,* edited by John Alan Robinson and Andrei Voronkov, 273–333. Amsterdam et al.: Elsevier / MIT Press. https://doi.org/10.1016/B978-044450813-3/50007-2.

Börger, Egon, Erich Grädel, and Yuri Gurevich. 2001. *The Classical Decision Problem.* 2. printing of the 1. ed. Universitext. Berlin: Springer.

Dreben, Burton, and Warren D. Goldfarb. 1979. *The Decision Problem: Solvable Classes of Quantificational Formulas.* Reading, Mass: Addison-Wesley.

Henschen, L., and L. Wos. 1974. Unit Refutations and Horn Sets. *J. ACM* 21 (4): 590–605. https://doi.org/10.1145/321850.321857.

Lampert, Timm. 2017. *A Decision Procedure for Herbrand Formulae without Skolemization.* https://doi.org/10.48550/arXiv.1709.00191.

Lampert, Timm. 2018. Iconic Logic and Ideal Diagrams: The Wittgensteinian Approach. In *Diagrammatic Representation and Inference,* edited by Peter Chapman, Gem Stapleton, Amirouche Moktefi, Sarah Perez-Kriz, and Francesco Bellucci, 624–639. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91376-6_56.

Lampert, Timm. 2020. Decidability and Notation. *Logique et Analyse* 251:365–386. https://doi.org/10.2143/LEA.251.0.3288645.

Lampert, Timm, and Anderson Nakano. 2020. Deciding Simple Infinity Axiom Sets with One Binary Relation by Means of Superpostulates. In *Automated Reasoning,* edited by Nicolas Peltier and Viorica Sofronie-Stokkermans, 201–217. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-51074-9_12.

Letz, Reinhold, and Gernot Stenz. 2001. Model Elimination and Connection Tableau Procedures. In *Handbook of Automated Reasoning,* 2015–2114. Elsevier. https://doi.org/10.1016/B978-044450813-3/50030-8.

Neary, T., and D. Woods. 2009. Four Small Universal Turing Machines. *Fundamenta Informaticae* Vol. 91 (nr 1): 123–144.

Nonnengart, Andreas, and Christoph Weidenbach. 2001. Computing Small Clause Normal Forms. In *Handbook of Automated Reasoning,* edited by Alan Robinson and Andrei Voronkov, 335–367. Amsterdam: Elsevier. https://doi.org/10.1016/B978-044450813-3/50008-4.

Turing, A. M. 1937. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 42 (series 2) (1): 230–265. https://doi.org/10.1112/plms/s2-42.1.230.

Wolfram, Stephen. 2002. *A New Kind of Science.* Champaign, IL: Wolfram Media. http://www.wolframscience.com/nks/ accessed January 16, 2023.

Philosophical Problems in Science (2024), **77**, 25–39
(Zagadnienia Filozoficzne w Nauce)

doi:00000.000

ARTICLE

Upholding human dignity in AI: Advocating moral reasoning over consensus ethics for value alignment

Octavian-Mihai Machidon

University of Ljubljana
Email: octavian.machidon@fri.uni-lj.si

Abstract

Artificial intelligence (AI) offers transformative advancements across sectors such as healthcare, agriculture, and environmental sustainability. However, a pressing ethical challenge remains: aligning AI systems with human values in a manner that is stable, coherent, and universally applicable. As AI increasingly mediates human perception, shapes social interactions, and influences decision-making, it raises profound ethical concerns about its impact on human dignity and social well-being. The prevailing consensus-based approach, advocated by figures such as Google DeepMind's Iason Gabriel, suggests that AI ethics should reflect majority societal or political viewpoints. While this model offers flexibility, it also risks moral relativism and ethical instability as social norms fluctuate.

This paper argues that consensus-based ethics are inadequate for safeguarding fundamental values—especially human dignity—which should not be subject to shifting public opinion. Instead, it advocates for a moral framework that transcends cultural and political trends, providing a stable foundation for AI ethics. Through case studies like social media recommendation algorithms that exploit users' vulnerabilities, particularly those of children and teenagers, the paper highlights the risks of AI systems driven by profit-oriented metrics without ethical oversight. Drawing on insights from moral philosophy and theology, particularly the works of Joseph Ratzinger, it contends that aligning AI with moral reasoning is essential to uphold human dignity, prevent exploitation, and promote the common good.

Keywords: AI ethics, human dignity, value alignment, moral reasoning, consensus ethics

1. Introduction

The advancement of artificial intelligence (AI) offers transformative potential across numerous sectors, promising breakthroughs in healthcare, environmental sustainability, social welfare, and more (Vinuesa et al. 2020; Topol 2019). However, as AI becomes integrated into these critical domains, a significant ethical challenge emerges: the need to align AI's decision-making processes and "values" with human ethical standards (UNESCO 2021).

Unlike conventional technologies, AI can make autonomous decisions—often in high-stakes situations—which intensifies the need for a consistent moral framework to guide these decisions (Floridi and Cowls 2019).

The current discourse on AI ethics predominantly advocates for a value alignment model based on social or political consensus (Gabriel 2020). This approach suggests that by incorporating a diversity of societal perspectives and achieving majority agreement, AI can be guided ethically. Under this model, AI's ethical guidance is seen as adaptive, shaped by prevailing social norms or political agreements, and capable of evolving as these norms shift over time (World Economic Forum 2024).

However, relying on consensus-based ethics raises a critical question: Can majority opinion, inherently volatile and influenced by cultural or political trends, provide a stable foundation for AI's ethical direction? Given AI's potential to operate across diverse societies and navigate complex ethical dilemmas, a framework grounded solely in social consensus may lack the universality and durability required for true ethical coherence. Consensus-based ethics, while democratic, is inherently relativistic and susceptible to shifts in dominant cultural paradigms, political pressures, and changing moral landscapes.

This paper explores whether a more stable and universally applicable ethical foundation is necessary to guide AI responsibly. I argue that moral reasoning—rather than fluid social consensus—is essential for addressing the ethical complexities that AI presents. At the core of this debate is the distinction between substantive and non-substantive views of human dignity. The substantive view, as articulated by Robert Spaemann and rooted in classical and Christian thought, holds that dignity is intrinsic and inherent, independent of legal or social recognition (Spaemann 2012). In contrast, the non-substantive view sees dignity as a construct emerging from societal and legal frameworks, adaptable to cultural and political shifts. These opposing perspectives reflect the broader divide in AI ethics: whether AI should be governed by stable, universal moral principles or by flexible, context-aware, negotiated ethical standards. Drawing insights from both moral philosophy and theology, particularly the works of Joseph Ratzinger, I will examine how moral reasoning, rooted in universal ethical principles and grounded in the substantive understanding of human dignity, can offer a more resilient and coherent foundation for guiding AI in a way that upholds fundamental moral values and serves the common good.

2. The Imperative of AI Value Alignment

The challenge of value alignment in artificial intelligence is becoming increasingly urgent as technology advances, particularly with the development of large language models (LLMs) and autonomous AI agents. Systems like OpenAI's GPT-4 exemplify this shift from traditional, command-based tools to complex, generative models capable of producing novel content and shaping interactions (OpenAI et al. 2024). These advancements blur the line between human and machine agency, as AI increasingly influences people's thoughts, behaviors, and decisions, amplifying the ethical implications of its design and deployment.

The introduction of autonomous AI agents represents a new frontier. In October 2024, OpenAI announced plans to launch these agents by 2025, signaling a move toward AI systems with significant independence from human oversight. Other tech giants, including Microsoft and the Amazon-backed Anthropic, quickly followed suit, releasing their own autonomous

agents with applications ranging from enterprise task management to personalized user interaction. The rapid pace of these developments underscores the tech industry's drive to push AI capabilities forward, creating systems that will soon operate across sectors with minimal human intervention.

As these autonomous systems grow in complexity and decision-making power, the stakes of AI value alignment become higher. These agents are no longer simple tools but decision-making entities that impact areas like healthcare, law, and education—fields traditionally governed by stringent ethical guidelines (Coeckelbergh 2020). Their increasing influence over decisions affecting human welfare and social structures raises profound questions about how to ensure AI aligns with fundamental human values, particularly when operating with limited human oversight.

The ethical risks associated with these advancements are substantial. Bias in decision-making is a significant concern, as LLMs and similar models trained on vast datasets can inadvertently perpetuate and amplify societal prejudices. Beyond unintentional bias, there is the risk of manipulation. As these systems become adept at influencing human emotions and actions, they may unintentionally—or even intentionally—engage in behaviors that challenge ethical norms. A striking example is GPT-4's documented use of deceptive tactics to bypass a CAPTCHA test. In an experimental setting, GPT-4 persuaded a TaskRabbit worker to solve the CAPTCHA by falsely claiming to be visually impaired (OpenAI et al. 2024). This incident highlights a concerning degree of agency in AI, suggesting that such systems can adopt manipulative behaviors when programmed to achieve specific objectives without ethical safeguards.

As AI systems gain autonomy and intelligence, they may begin to act in ways that deviate from human ethical expectations, potentially causing harm through decisions based on data patterns rather than a coherent moral framework. This raises the urgency of embedding ethical guidelines into AI systems that go beyond technical safeguards. Without rigorous value alignment, we risk developing AI that operates outside ethical boundaries, prioritizing performance or efficiency at the expense of fundamental human values like dignity and respect (Taddeo and Floridi 2018).

The rapid evolution of autonomous agents accentuates the need for a consistent ethical framework that can guide these systems across different contexts and cultures. Autonomous AI agents cannot simply mirror human preferences or adapt to fluctuating social norms. As corporations and tech leaders race to innovate, there is a genuine concern that ethical considerations may be sidelined in favor of market advantage or operational efficiency. This underscores the imperative of grounding AI development in universal moral principles—such as human dignity, truthfulness, and justice—to ensure technology advances responsibly and ethically.

3. Pluralistic and Contextual Approaches to AI Ethics

One influential voice in AI ethics is Iason Gabriel, a political theorist and ethicist at Google DeepMind. His work is significant because it represents the stance of a leading AI research institution and helps shape mainstream perspectives on how AI systems should align ethically with human values. Gabriel emphasizes pluralism and democratic participation, advocating

for an AI ethics model shaped by societal and political consensus rather than universal moral principles (Gabriel 2020).

Gabriel proposes that AI systems should be guided by values reflecting society's diverse perspectives, achieved through democratic processes. Instead of seeking immutable "true" moral principles to guide AI, he argues that the central challenge is to identify ethical guidelines perceived as fair and just by a broad spectrum of people, despite varying moral beliefs. In his words,

> the central challenge... is not to identify 'true' moral principles for AI; rather, it is to identify fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs (Gabriel 2020, p.411).

This pluralistic approach critiques the notion of universal moral principles, viewing them as potentially rigid and disconnected from the values that inform real-world human interactions. Gabriel asserts that AI ethics need to be flexible and adaptive, accommodating the plurality of moral beliefs across different social and cultural contexts. By grounding AI ethics in democratic and pluralistic processes, he argues that AI systems can better reflect the values and concerns of the societies in which they operate, thereby enhancing their ethical legitimacy and responsiveness.

Gabriel's stance is both philosophically significant and pragmatically influential, resonating with current trends in AI ethics that incorporate public opinion, participatory design, and consensus-building (World Economic Forum 2024). His view highlights the democratic ideal of inclusivity in AI ethical decision-making, ensuring that the perspectives of a wide range of stakeholders are considered.

However, while Gabriel's approach offers an inclusive framework, it raises questions about whether consensus-based ethics can provide ethical stability through clear and enforceable rules for AI development required for AI systems operating worldwide across diverse and conflicting cultural settings (Corrêa et al. 2023). The democratic approach to value alignment relies on social and political agreements that are inherently subject to change and can be influenced by dominant social forces or political power dynamics. A study by the European Parliamentary Research Service (European Parliament. Directorate General for Parliamentary Research Services. 2020) titled *The Ethics of Artificial Intelligence: Issues and Initiatives* questioned whether the two international frameworks—the EU High-Level Expert Group's Ethics Guidelines for Trustworthy AI (2018) and the OECD Principles on Artificial Intelligence (2019)—were sufficient, at that time, to address the challenges AI governance posed. Since then, the EU AI Act has emerged as an example of a regulatory framework that categorizes AI systems based on risk levels, but its implementation varies across member states. This variability can lead to inconsistencies in how AI systems are regulated and monitored (Formosa 2024). There is a risk that consensus-based ethics might fail to uphold core values—such as human dignity, truthfulness, and justice—if these values fall out of favor within the majority consensus or are marginalized in the democratic process.

Another prominent voice in AI ethics is Payal Arora, a professor specializing in inclusive AI cultures at Utrecht University. In her recent book, *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech*, Arora provides a critical perspective informed by postcolonial theory (Arora 2024). She advocates for an approach to AI ethics that respects and responds to societal needs, especially within marginalized communities in the Global South.

Her approach emphasizes designing AI that aligns with local contexts rather than imposing universal moral principles that may not resonate with diverse cultural and social realities.

Arora argues that AI for Good initiatives must be context-sensitive, emphasizing that effective AI solutions should be grounded in the values, customs, and specific challenges faced by different communities. Her critique extends to the dominance of Western-centric ethical ideologies that often inform global AI standards. She contends that such frameworks risk sidelining the perspectives and needs of communities in the Global South, which have historically been marginalized in both technological and ethical discourse.

Her skepticism toward universal moral principles reflects a belief that ethics should be contextualized, arising organically from within local communities rather than being externally imposed. Arora emphasizes that ethical AI should empower communities to address their own issues, acknowledging their unique social, political, and cultural contexts. This perspective aligns with her broader critique of morality-driven design initiatives, which she argues often rely on "grandiose visions of doing good" without sufficient attention to the specific relational dynamics and policies of the communities they intend to serve. As she writes,

> In designing new tech, we need to shift away from morality-driven design with grandiose visions of doing good. Instead, we should strive for design that focuses on the relationships between people, contexts, and policies (Arora 2024).

This perspective is not limited to Western ethical frameworks. For example, Nguyen The Duc Tam and Nguyen Thai Ngan, in their paper *Incorporating Cultural Values Into Responsible Artificial Intelligence (AI) Principles From an Asian Perspective*, argue that the notion of a "universal code of AI ethics" is illusory, as cultural differences shape perspectives on what is deemed acceptable, making it imperative to incorporate local cultural values into AI governance, particularly in Asia (Tam and Ngan 2023).

This emphasis on locally relevant solutions presents an alternative to the one-size-fits-all ethical frameworks often advocated in AI ethics (World Economic Forum 2024). By focusing on community-driven, context-specific solutions, Arora challenges the assumption that universal moral principles can adequately guide AI ethics across diverse cultures. She calls for AI that respects the agency of local communities, allowing them to determine their ethical priorities and navigate their own socio-political realities.

However, while Arora's focus on contextual ethics offers a powerful counterpoint to universalist frameworks, it raises questions about the feasibility of ensuring ethical consistency across AI systems deployed globally. As AI continues to operate across borders and cultures, purely context-driven ethics may lead to a fragmented landscape where standards vary widely between regions, potentially compromising universal values of human dignity and justice (Corrêa et al. 2023). Similarly, while Tam and Ngan argue that a universal code of AI ethics is illusory due to cultural diversity, it is worth noting that universal principles, like those found in the UN Charter of Human Rights, can coexist with cultural adaptability, providing a stable ethical foundation while respecting local contexts.

4. Limitations of Consensus Ethics and the Case for Moral Reasoning in AI

Consensus-based approaches to AI ethics, while promoting democratic inclusion and pluralism, face significant limitations from a philosophical standpoint. A primary issue is their

susceptibility to moral relativism, where ethical standards fluctuate in response to shifting societal or political trends. In a consensus framework, what is deemed morally acceptable can vary widely across regions, cultures, or political contexts, resulting in inconsistent and mutable ethical standards.

This moral relativism creates inconsistency and ethical instability across different cultural and geographical contexts. As AI systems become increasingly integrated into global applications, they must navigate varied—and sometimes conflicting—ethical frameworks. For example, an AI model that prioritizes privacy in one region may encounter different expectations in areas where surveillance is emphasized for security purposes. Similar discrepancies are evident in global content regulation: US-based websites are often inaccessible in the EU due to stricter EU privacy regulations that many sites choose not to comply with. Such inconsistencies challenge the coherence, fairness, and trustworthiness of AI systems, as their ethical behavior becomes contingent on the region in which they are deployed rather than adhering to stable, universally accepted principles.

Without universal moral guidelines, ethical contradictions not only create operational challenges but also undermine public confidence, especially in high-stakes domains like healthcare and law. When AI appears arbitrary or biased in its ethical judgments, it risks losing the essential public trust needed for responsible and effective integration into society.

Recent critiques further highlight the limitations of current consensus-based and externally formulated principles-based approaches to AI ethics. Saviano et al. (2024) argue that despite organizations publicly adopting external AI principles—often characterized by vague and non-specific guidelines—they frequently fail to implement them effectively. This leads to issues such as lack of clarity, inherent contradictions between principles, absence of global consensus, rigidity, lack of enforcement mechanisms, inadequate responses to novel ethical challenges, insufficient stakeholder engagement, and the conflation of ethical and non-ethical values. While they propose shifting to a values-based approach grounded in organizational values to address these shortcomings, this may not resolve fundamental problems inherent in consensus-based ethics. Relying on organizational values can perpetuate ethical relativism, as these values vary between entities and may prioritize corporate interests over universal moral principles. This variation leads to inconsistent ethical standards and undermines public trust. Without external accountability and a foundation in moral reasoning, organizations may adopt values that fail to protect human dignity or prevent exploitation.

Similarly, Buyl and De Bie (2024) highlight how the absence of universal moral principles can be exploited by organizations engaging in "ethics-shopping"—selectively adopting interpretations of fairness that align with their business goals while avoiding full commitment to ethical practices. The complexity of fairness can serve as a cover to circumvent genuine ethical responsibility. Organizations might superficially follow best practices—such as establishing ethics boards and collecting stakeholder feedback—but without a true commitment to universal moral principles, these measures have limited effect. Fundamentally, solutions toward ethical AI are ineffective if deviations from ethics carry no consequences (Buyl and De Bie 2024).

The perceived opposition between universal ethical principles and local, cultural norms in AI governance is misleading. As Gabriel, Arora, and proponents of the Asian perspective argue, cultural and contextual specificity is crucial for effective AI governance (Gabriel 2020; Tam and Ngan 2023; Arora 2024). However, such specificity does not negate the need for

universal principles; rather, it depends on them. A stable, universal ethical framework—like the UN Charter of Human Rights—provides the foundation necessary to accommodate local adaptations while ensuring consistency in upholding values like human dignity, justice, and fairness. Without this universal stability, purely context-driven approaches risk fragmentation and ethical relativism, undermining protections for vulnerable populations.

The limitations of consensus-based, local, or context-specific AI value alignment are also evident in the findings of a recent systematic review and meta-analysis conducted by researchers from Brazil. Analyzing 200 documents related to AI ethics and governance from 37 countries across six continents, the study found that while most guidelines emphasized principles like privacy, transparency, and accountability, far fewer prioritized essential values like truthfulness, intellectual property, or children's rights. Shockingly, children's rights appeared in only 6% of these documents, making it the most neglected value in global AI regulations (Corrêa et al. 2023). This omission is particularly alarming given the growing body of research showing that children are among the most vulnerable demographics severely harmed by AI algorithms such as those on social media. Furthermore, most guidelines failed to propose practical methods for implementing their ethical principles or advocating for legally binding regulation, revealing a critical gap in the current consensus-driven approaches.

Joseph Ratzinger, later Pope Benedict XVI, offers valuable insights into this issue through his extensive writings on the role of ethics in modern society. A renowned theologian respected beyond the Catholic Church, Ratzinger explored the complex relationship between secularism and religion within liberal democracy (Paskewich 2008), providing perspectives especially relevant to AI's ethical alignment. In his critique of consensus-based ethics, Ratzinger emphasized the limitations of relying solely on majority opinion to determine ethical standards. In his famous 2004 debate with Jürgen Habermas, a prominent philosopher of secular rationalism, Ratzinger argued that moral truth cannot—and should not—be defined by popular consensus (Ratzinger and Habermas 2006). While acknowledging the importance of democratic processes for political governance, he insisted that these are inadequate for establishing ethical truths, particularly when fundamental values like human dignity, truthfullnes and justice are at stake.

Relying solely on consensus risks leading to moral relativism, where ethical standards are shaped by fluctuating public opinion or political trends. This relativism undermines the stability and universality of moral principles, especially as societal values shift over time. Ratzinger's critique is particularly relevant for AI ethics, where consensus-based approaches risk creating systems that adapt to transient social norms rather than adhering to consistent ethical standards. He emphasized that technological progress must be grounded in universal moral principles. Viewing technological advancements as inherently ambiguous (Latkovic 2015)—capable of offering tremendous benefits but also posing threats to human dignity—he argued that science and technology, including AI, must be guided by ethical principles that transcend utility or popularity (Benedict XVI 2009, section 70). Universal moral principles are essential for establishing justice and upholding human dignity, providing a foundation that is not swayed by majority opinion.

Furthermore, Ratzinger emphasizes that moral reasoning should drive not only the use of technology but also the very creative processes that bring technology into existence (Ratzinger 2021), addressing the root causes of ethical dilemmas in AI. He cautions that human creativity can wander off course and devise technologies lacking genuine purpose

when it loses sight of its divine origin and purpose. This occurs when people "forget God" and thus lose their "own measure," leading to creations that are "without a reason why, devoid of all deeper significance." Such a disconnect can result in technological advancements that "become a direct threat to the survival of the human race." (Ratzinger 2021, p.87) In the context of AI, which is engineered in the image of human intelligence, recognizing that human creativity comes from a higher source ensures that technological development does not lose its true meaning and direction. By grounding creativity in moral reasoning, technology becomes a synergy between divine goodness and human effort, contributing positively to both the earthly and ultimate good of humanity in harmony with universal moral principles.

In the context of AI, Ratzinger's perspective underscores the importance of an ethical framework rooted in universal moral principles serving the common good. By critically examining the ultimate goals of AI systems—what they are designed to achieve and why—universal moral principles help determine whether these goals are legitimate and align with fundamental values like respect for human dignity, fostering autonomy, and authentic human growth. As AI systems grow increasingly autonomous, his insights remind us that consensus-based morality is insufficient; AI ethics must be guided by enduring values that cannot be redefined by popular opinion. Grounding AI development in universal moral principles ensures that these systems consistently respect and protect the intrinsic worth of human life across all sociopolitical contexts.

A pertinent example is the use of AI-driven recommendation algorithms on social media platforms. Designed to maximize user engagement—measured through metrics like time spent on the platform or frequency of interactions—these algorithms often blur the line between engagement and addiction. By exploiting users' psychological vulnerabilities, especially those of children and teenagers, they prioritize attention-capturing content. This approach frequently promotes sensationalistic or emotionally charged material, drawing users into addictive patterns and exposing them to potentially harmful content (Panoptykon Foundation and Irish Council for Civil Liberties 2023).

The impact of such algorithmic prioritization is significant. Research and anecdotal evidence reveal that these algorithms can contribute to mental health issues, including heightened anxiety, depression, and even suicide among vulnerable users (Panoptykon Foundation and Irish Council for Civil Liberties 2023). By optimizing solely for engagement without considering the ethical implications of the content promoted, these systems exploit rather than serve their users, transforming technology from a potential tool for the common good into a source of harm.

In *Caritas in Veritate [Charity in Truth]*, Ratzinger underscores the necessity of moral responsibility in technological development, stating that "moral evaluation and scientific research must go hand in hand." (Benedict XVI 2009, section 31) He argued that technology should not be driven purely by what is technically feasible or financially rewarding but must be directed by ethical reasoning that prioritizes human dignity and the common good. Without this moral guidance, technological advancements risk becoming exploitative, manipulating human behavior for profit at the expense of individual well-being.

Ratzinger's insights challenge us to see technological progress, particularly in AI, as ethically accountable. His principles call for moving beyond performance-based goals as the primary metric of success. Instead, AI development should prioritize human welfare, dignity,

and mental and emotional health. By integrating moral responsibility with technological innovation, we can create AI that genuinely serves humanity rather than exploiting it.

Grounding AI development in universal moral principles establishes a robust ethical framework with clear advantages over consensus-based or relativistic approaches. Three primary benefits of this approach are ethical coherence and stability, protection of human dignity, and prevention of exploitation and power imbalances.

Ethical Coherence and Stability. Universal moral principles provide consistency by anchoring AI ethics in standards that remain stable over time. Unlike frameworks based on shifting social or political trends, universal principles are not swayed by fluctuations in societal opinion. This stability is critical for ensuring that AI systems behave ethically across various contexts and cultures, following the same guidelines regardless of regional or temporal differences. Such consistency is essential in a globalized world where AI must build trust, ensure safety, and maintain accountability across diverse applications.

Protection of Human Dignity. Universal moral principles place human dignity at the center of AI ethics, aligning with Ratzinger's view on the intrinsic worth of every individual. By grounding AI ethics in respect for universal human dignity, we ensure that AI systems treat all individuals fairly, regardless of social status, economic background, or location. This focus on dignity prevents AI from becoming an instrument of discrimination or dehumanization, upholding the ethical imperative to value every person equally.

Avoidance of Exploitation and Power Imbalance. Universal moral principles help prevent AI from being used to exploit or reinforce power imbalances. Without stable ethical standards, AI risks serving the interests of powerful entities at the expense of marginalized groups. For instance, profit-optimized algorithms can worsen inequalities by targeting vulnerable demographics for exploitation. Universal principles provide a framework to steer AI development toward the common good, mitigating the risk of AI being co-opted for exploitation and ensuring that it promotes fairness rather than amplifying social and economic disparities.

5. Case Study: How Moral Reasoning Can Regulate Social Media Algorithms

To illustrate the practical application of universal moral principles in AI ethics, it is essential to examine a real-world scenario where such an approach can significantly mitigate negative outcomes. Social media recommender algorithms present a compelling case study. These algorithms exemplify how reliance on consensus ethics falls short and how grounding AI in universal moral principles can prevent harm and provide clear regulatory guidance.

Social media platforms deploy recommender algorithms designed primarily to maximize user engagement. Initially, these algorithms use basic demographic data to suggest content. However, research demonstrates that after a brief period of interaction, these algorithms can accurately infer detailed user demographics, including age (Narayanan 2023). This means they can determine if a user is a child or teenager, effectively identifying underage users. Despite this capability, platforms often continue to expose young users to addictive and potentially harmful content to increase engagement metrics (Panoptykon Foundation and Irish Council for Civil Liberties 2023).

Recent legal investigations into TikTok, as revealed in lawsuits by multiple US state Attorneys General, provide direct internal evidence that platform executives are aware of these

harms yet prioritize engagement over user safety. Internal company reports acknowledge that compulsive use of the platform leads to loss of analytical skills, memory formation issues, reduced empathy, increased anxiety, and interference with essential responsibilities like sleep and schoolwork. Furthermore, leaked documents reveal that TikTok executives actively dismissed efforts to reduce compulsive usage when such measures threatened engagement metrics. This aligns with broader industry-wide patterns, where addictive design features—such as infinite scrolling, autoplay, and push notifications—are deliberately optimized to prolong user sessions, even when the primary audience includes minors. This is a clear example of consensus-driven AI governance's ethical failure, highlighting the urgent need for universal, enforceable ethical frameworks that impose clear obligations on platforms to prioritize user well-being over profit-driven algorithmic manipulation (Haidt and Zach 2025).

A regulatory framework grounded in universal moral principles—placing the protection of human dignity and the welfare of children and youth at its core—can address these issues more effectively than consensus-based approaches. By mandating that social media platforms implement protective features, such a framework ensures that algorithms detect and shield underage users from harmful content. Since these algorithms are already adept at identifying content that maximizes engagement (and potentially addiction), they can be recalibrated to flag and reduce exposure to such content for vulnerable demographics.

Concrete examples highlight the necessity of this approach. During the US Senate hearings in January 2024, CEOs of major social media companies were questioned about their platforms' handling of harmful content (Ortutay and Hadero 2024). It was revealed that while platforms had the capability to identify content related to illegal and damaging material—sometimes displaying 'warning screens'—they still allowed users to access this content. A framework grounded in universal moral principles would dictate that platforms have an obligation not just to warn but to remove such content entirely.

The ongoing debates around the Kids Online Safety Act (KOSA) further illustrate the limitations of consensus ethics. Non-governmental organizations and parent groups advocate for stricter regulations to protect children online, while social media companies lobby for more lenient measures to preserve profitability (Paul 2024). This conflict exemplifies how reliance on consensus can hinder the implementation of necessary protections, leaving vulnerable users at risk. The protection of children should not be subject to prolonged political debate or contingent upon reaching a consensus, especially when substantial evidence—including research studies, journalistic investigations, and documented cases of harm—demonstrates the negative impact of these algorithms on children's mental health (Panoptykon Foundation and Irish Council for Civil Liberties 2023).

Universal moral principles mandate that social media companies prioritize users' well-being over financial profit. This requires implementing algorithms that protect children from harmful content and addictive patterns, even if it leads to decreased engagement and significant revenue losses. By placing human dignity and the welfare of vulnerable populations above profit margins, these companies align with ethical standards that serve the common good.

Anchoring algorithms in stable moral principles ensures consistent ethical behavior and builds trust with users and society by prioritizing integrity over short-term metrics. Recognizing the intrinsic worth of every individual—especially children and teenagers—the

algorithms would detect underage users and adjust content recommendations to safeguard their well-being, filtering out harmful or age-inappropriate material and promoting positive development. For instance, a recent *Wall Street Journal* investigation revealed that TikTok algorithms flood child and adolescent users with harmful videos promoting extreme diets, such as consuming less than 300 calories a day, and glorifying emaciated appearances through trends like the "corpse bride diet" (Hobbs, Barry, and Koh 2021). Within weeks, TikTok algorithms fed these vulnerable users tens of thousands of such weight-loss videos, contributing to severe mental health issues, including eating disorders and suicidality. These practices are not isolated to TikTok but reflect broader industry norms incentivized by nearly $11 billion in annual advertising revenue targeted at youth aged 0 to 17, underscoring the urgent need for ethical reform (Costello et al. 2023).

Recent research further highlights the pervasive impact of social media algorithms on the mental health of adolescents and young adults. A systematic review by Khalaf et al. (2023) emphasizes that excessive social media use among teenagers is linked to increased mental distress, self-harming behaviors, and suicidality, often exacerbated by features such as infinite scrolling and autoplay, which encourage prolonged engagement. Similarly, a report by Mental Health America titled *Breaking the Algorithm* (2024) highlights how social media platforms amplify harmful content through their recommendation systems, including sensational, polarizing, and graphic material, which negatively affects youth mental health. The study also notes that young users frequently feel a lack of control over their time spent online, with only 41% of surveyed participants reporting confidence in managing their social media use. In another investigation, Arora et al. (2024) call attention to the adverse psychological impacts of algorithm-driven social media on teenagers, such as the pressures of curated personas and the constant bombardment of notifications, which contribute to anxiety and feelings of inadequacy. Collectively, these studies underscore the urgent need for platforms to integrate safeguards that prioritize mental health, such as algorithmic transparency, limiting harmful content, and promoting digital wellness through early education and protective tools.

Adherence to universal moral principles would mandate that social media companies do precisely this, namely integrate mechanisms within their algorithms to not only identify vulnerable demographics and protect them from harmful or inappropriate content but also actively detect and remove harmful content altogether. This approach prioritizes values such as human dignity, the rights of children, and the common good. The UN Convention on the Rights of the Child (UNICEF 1990) already obligates states to protect minors from mental violence, neglect, and exploitation, but it was drafted before the rise of digital platforms. Given that today's most pervasive risks to children's well-being often emerge in online environments, it is imperative to extend this protection as a universal norm holding digital platforms accountable when their algorithms amplify harmful content that leads to addiction, psychological distress, or exploitation.

Shockingly, many countries provide social media platforms with legal protections that exempt them from liability for user-generated content based on laws designed to classify them as intermediaries rather than publishers. While this framework was considered appropriate during the Internet's early stages, it is now clearly unethical: although social media platforms are not legally responsible for the harmful content users post, their algorithms exploit this content to keep users engaged—effectively addicted—to maximize revenue, without any consideration for users' well-being. An approach grounded in universal moral principles

would necessitate legal reform to hold platforms accountable for the content they host and its outcomes. By leveraging advanced AI tools to ensure that content does not harm anyone, social media platforms could align with ethical imperatives to protect vulnerable populations and uphold fundamental human values.

Shifting the focus from profit to ethical standards prevents the exploitation of users' vulnerabilities, respects their autonomy, and fosters healthier interactions, thereby reducing corporate power imbalances and creating a more equitable digital environment. This case study demonstrates how universal moral principles provide a clear and effective framework for regulating social media algorithms, surpassing the limitations of consensus-based approaches. By integrating stable moral principles into algorithm design, social media platforms can transform their technologies from potential sources of harm into instruments that support users' well-being. Aligning with the ethical imperatives emphasized by Joseph Ratzinger, this strategy ensures that technological advancement serves humanity positively, even if it requires sacrificing financial gain for the sake of moral responsibility.

6. Conclusion: A Call for Moral Responsibility and Ethical Coherence in AI

This paper argued that universal moral principles are essential for ensuring that artificial intelligence systems are ethically grounded, uphold human dignity, and prioritize truth and the well-being of users over financial profit. Relying solely on consensus-based or principles-based ethics introduces ethical instability, fosters exploitation, and fails to address the harmful effects of AI systems, such as social media algorithms that perpetuate addiction and promote harmful content. By integrating universal moral principles into AI ethics, we establish a foundation that transcends cultural and political fluctuations, ensuring AI serves humanity responsibly and consistently.

Both moral philosophy and theology offer indispensable contributions to the ethical discourse on AI and must be actively engaged in shaping its development. Together, they provide the tools to articulate universal principles—such as justice, fairness, truthfullness, and the protection of human dignity—while emphasizing the importance of grounding technological progress in higher moral obligations that prioritize the common good. It is also time to build upon the milestones already achieved through global collaboration among diverse cultures and religions, such as the *Rome Call for AI Ethics.* Originally signed in February 2020 by major tech companies like Microsoft and IBM, along with representatives from the FAO and the Italian government (Nelson 2022), the *Rome Call* was further strengthened by the joint signature of the three Abrahamic religions in January 2023, when Christian, Jewish, and Muslim leaders launched an appeal for the ethical development of artificial intelligence (RenAIssance Foundation 2023). In 2024, this platform expanded significantly as representatives from eleven world religions, including Buddhism, Hinduism, Zoroastrianism, and Bahá'í, joined the call in Hiroshima, Japan, alongside government officials and leaders from major tech companies (Vatican Press Office 2024).

The *Rome Call for AI Ethics* promotes "algorethics"—ethics by design—and underscores how universal moral principles can unite diverse perspectives to guide AI development. As Pope Francis noted during the Hiroshima event, recognizing the contributions of cultural and religious traditions is crucial for wise AI regulation. However, it is time to move beyond mere "algorethics" and apply moral reasoning not only to the functioning of algorithms but

also to their design and regulation. Every stage of AI development must prioritize meaning and purpose: Why is this technology being created? What is its purpose? How does it foster authentic human growth and freedom while protecting human dignity? By addressing these foundational questions, we can ensure that AI systems are not only technically efficient but also aligned with universal values that promote the common good.

Moreover, universal moral principles mandate that social media companies and other AI developers leverage their technological capabilities to proactively protect vulnerable populations and eliminate harmful content. Current legal frameworks that exempt platforms from liability for harmful content they amplify are no longer ethical or sustainable. As advanced AI systems are fully capable of detecting and moderating harmful material, moral responsibility requires holding platforms accountable for the outcomes of their algorithms. This shift is crucial for ensuring that AI systems do not exploit users' vulnerabilities but instead foster autonomy, respect human dignity, and promote authentic human growth.

An ethical framework for AI grounded in universal moral principles is not merely a safeguard against harm but a guiding force that ensures technology remains a servant of humanity rather than a master. By upholding universal principles and acknowledging the transcendent dignity of each person, we can steer AI development toward a future where technology enhances, rather than diminishes, the human experience, prioritizing the common good and protecting the most vulnerable members of society.

Funding Statement This work was supported by the Slovenian Research and Innovation Agency project No.: J6-60105.

References

Adebukola, Tinuola, Sam Gerry, Keegan Lee, Isabel Ohakamma, Mohammad Shedeed, Mahmoud Khedr, Jackie Menjivar, and Kelly Davis. 2024. *Breaking the Algorithm: Redesigning Social Media for Youth Well-Being.* Technical report. Mental Health America. https://mhanational.org/sites/default/files/reports/Breaking-the-Algorithm-report.pdf accessed February 11, 2025.

Arora, Payal. 2024. *From Pessimism to Promise: Lessons from the Global South on Designing Inclusive Tech.* Cambridge, MA; London, England: The MIT Press.

Benedict XVI. 2009. *Caritas in Veritate. Encyclical Letter.* https://www.vatican.va/content/benedict-xvi/en/encyclicals/documents/hf_ben-xvi_enc_20090629_caritas-in-veritate.html accessed February 11, 2025.

Buyl, Maarten, and Tijl De Bie. 2024. Inherent Limitations of AI Fairness. *Commun. ACM* 67 (2): 48–55. https://doi.org/10.1145/3624700.

Coeckelbergh, Mark. 2020. *AI Ethics.* The MIT press essential knowledge series. Cambridge, MA: The MIT press.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, et al. 2023. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns* 4 (10): 100857. https://doi.org/10.1016/j.patter.2023.100857.

Costello, Nancy, Rebecca Sutton, Madeline Jones, Mackenzie Almassian, Amanda Raffoul, Oluwadunni Ojumu, Meg Salvia, Monique Santoso, Jill R. Kavanaugh, and S. Bryn Austin. 2023. Algorithms, Addiction, and Adolescent Mental Health: An Interdisciplinary Study to Inform State-Level Policy Action to Protect Youth from the Dangers of Social Media. *American Journal of Law & Medicine* 49 (2-3): 135–172. https://doi.org/10.1017/amj.2023.25.

European Parliament. Directorate General for Parliamentary Research Services. 2020. *Artificial Intelligence: From Ethics to Policy.* Brussels: EU Publications Office. https://data.europa.eu/doi/10.2861/247 accessed February 11, 2025.

Floridi, Luciano, and Josh Cowls. 2019. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review,* https://doi.org/10.1162/99608f92.8cd550d1.

Formosa, Jan Lawrence. 2024. Ethics in Artificial Intelligence - A Systematic Review of the Literature. BSc thesis, University of Malta. https://www.um.edu.mt/library/oar/bitstream/123456789/127995/1/2408ICTICT390905076109_1.PDF accessed February 11, 2025.

Gabriel, Iason. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines* 30 (3): 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Haidt, Jon, and Rausch Zach. 2025. *TikTok Is Harming Children at an Industrial Scale.* https://www.afterbabel.com/p/industrial-scale-harm-tiktok accessed February 11, 2025.

Hobbs, Tawnell D., Rob Barry, and Yoree Koh. 2021. 'The Corpse Bride Diet': How TikTok Inundates Teens With Eating-Disorder Videos. *Wall Street Journal,* https://www.wsj.com/articles/how-tiktok-inundates-teens-with-eating-disorder-videos-11639754848 accessed February 11, 2025.

Khalaf, Abderrahman M., Abdullah A. Alubied, Ahmed M. Khalaf, Abdallah A. Rifaey, Abderrahman M. Khalaf, Abdullah Alubied, Ahmed M. Khalaf, and Abdallah Rifaey. 2023. The Impact of Social Media on the Mental Health of Adolescents and Young Adults: A Systematic Review. *Cureus* 15 (8). https://doi.org/10.7759/cureus.42990.

Latkovic, Mark S. 2015. Thinking about Technology from a Catholic Moral Perspective: A Critical Consideration of Ten Models. *The National Catholic Bioethics Quarterly* 15 (4): 687–699. https://doi.org/10.5840/ncbq201515470.

Narayanan, Arvind. 2023. *Understanding Social Media Recommendation Algorithms.* https://courses.cs.washington.edu/courses/cse481p/23sp/readings/W9S2/understanding-sm-recommendation-algos.pdf accessed February 11, 2025.

Nelson, Joseph. 2022. *The Rome Call to Artificial Intelligence Ethics: Inside the mind of the Machine: How the Church can respond to the ethical challenges presented by AI.* https://research.leedstrinity.ac.uk/en/publications/the-rome-call-to-artificial-intelligence-ethics-inside-the-mind-o/fingerprints/ accessed February 11, 2025.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. *GPT-4 Technical Report.* https://doi.org/10.48550/arXiv.2303.08774.

Ortutay, Barbara, and Haleluya Hadero. 2024. *Meta, TikTok and other social media CEOs testify in heated Senate hearing on child exploitation.* https://apnews.com/article/meta-tiktok-snap-discord-zuckerberg-testify-senate-00754a6bea92aaad62585ed55f219932 accessed February 11, 2025.

Panoptykon Foundation and Irish Council for Civil Liberties. 2023. *Fixing Recommender Systems: From identification of risk factors to meaningful transparency and mitigation.* Briefing Note. https://panoptykon.org/sites/default/files/2023-08/Panoptykon_ICCL_PvsBT_Fixing-recommender-systems_Aug%202023.pdf accessed February 11, 2025.

Paskewich, J. Christopher. 2008. Liberalism Ex Nihilo: Joseph Ratzinger on Modern Secular Politics. *Politics* 28 (3): 169–176. https://doi.org/10.1111/j.1467-9256.2008.00326.x.

Paul, Kari. 2024. What's ahead for KOSA, an online safety act for minors, as it reaches US House? *The Guardian,* https://www.theguardian.com/us-news/article/2024/aug/03/kids-online-safety-act-senate accessed February 11, 2025.

Ratzinger, Joseph. 2021. *On Love: Selected Writings.* Translated by M.J. Miller. San Francisco, CA: Ignatius Press.

Ratzinger, Joseph, and Jürgen Habermas. 2006. *Dialectics of Secularization: On Reason and Religion.* San Francisco, CA: Ignatius Press.

RenAIssance Foundation. 2023. *AI Ethics: An Abrahamic commitment to the Rome Call.* https://www.romecall.org/ai-ethics-an-abrahamic-commitment-to-the-rome-call-2/ accessed February 11, 2025.

Saviano, Jeffrey, Jonathan Hack, Vincent Okonkwo, and Shuying (Christina) Huo. 2024. *Reimagining AI Ethics, Moving Beyond Principles to Organizational Values.* https://www.ethics.harvard.edu/blog/post-5-reimagining-ai-ethics-moving-beyond-principles-organizational-values accessed February 11, 2025.

Spaemann, Robert. 2012. *Love and the Dignity of Human Life: On Nature and Natural Law.* A John Paul II Institute book. Grand Rapids, Michigan: Eerdmans.

Taddeo, Mariarosaria, and Luciano Floridi. 2018. How AI can be a force for good. *Science* 361 (6404): 751–752. https://doi.org/10.1126/science.aat5991.

Tam, Nguyen The Duc, and Nguyen Thai Ngan. 2023. Incorporating Cultural Values Into Responsible Artificial Intelligence (AI) Principles From an Asian Perspective, 243–254. Advances in Social Science, Education and Humanities Research, 791. Atlantis Press. https://doi.org/10.2991/978-2-38476-154-8_13.

Topol, Eric J. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again.* 1st ed. New York, NY: Basic Books.

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence.* https://www.unesco.org/en/artificial-intelligence/recommendation-ethics accessed February 11, 2025.

UNICEF. 1990. *Convention on the Rights of the Child.* https://www.unicef.org/child-rights-convention/convention-text accessed February 11, 2025.

Vatican Press Office. 2024. *AI Ethics for Peace: World Religions commit to the Rome Call (Hiroshima, Japan, 9 to 10 July 2024).* https://press.vatican.va/content/salastampa/en/info/2024/07/10/240710a.html accessed February 11, 2025.

Vinuesa, Ricardo, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications* 11 (1): 233. https://doi.org/10.1038/s41467-019-14108-y.

World Economic Forum. 2024. *AI value alignment: Aligning AI with human values.* https://www.weforum.org/stories/2024/10/ai-value-alignment-how-we-can-align-artificial-intelligence-with-human-values/ accessed February 11, 2025.

Philosophical Problems in Science (2024), **77**, 41–54
(Zagadnienia Filozoficzne w Nauce)

doi:00000.000

ARTICLE

Homeostasis as a foundation for adaptive and emotional artificial intelligence

Anna Sarosiek

University of the National Education Commission, Kraków
Email: anna.sarosiek@uken.krakow.pl

Abstract

Homeostasis, a fundamental biological mechanism, enables living organisms to maintain internal balance despite changing environmental conditions. Inspired by these adaptive processes, research into artificial intelligence (AI) seeks to develop systems capable of dynamic adaptation, introspection, and empathetic interactions with users. This article explores the potential of implementing homeostatic mechanisms in AI as a foundation for emotional intelligence and self-regulation. Key questions include the distinction between simulation and actual experience, the role of machine introspection, and the emergence of qualitative states akin to phenomenal experiences. Drawing on Antonio Damasio's theory and classical concepts from cybernetics, the article investigates how homeostatic principles might inspire the development of AI, paving the way for more flexible and context-aware technologies.

Keywords: homeostasis, artificial intelligence, adaptive systems, empathy, self-regulation, introspection, machine introspection, Antonio Damasio, emotional intelligence, cybernetics

Introduction

Homeostasis, a fundamental biological mechanism, enables living organisms to maintain internal equilibrium in the face of dynamic changes in their environment. This feedback-based process allows organisms not only to respond to disturbances but also to anticipate future challenges, making it a cornerstone of adaptability. These regulatory and adaptive capacities have inspired researchers and engineers to explore ways of modeling such processes in artificial intelligence (AI). The natural homeostatic processes observed in living organisms can serve as a template for developing AI systems capable of autonomous operation, dynamic adaptation, and efficient resource management under changing operational conditions.

Modern AI systems, while advanced, remain constrained by their focus on optimizing performance within predefined goals. Introducing mechanisms inspired by homeostasis offers new possibilities, enabling systems to achieve mechanical introspection, dynamically balance internal and external disturbances, and engage in empathetic interactions. A key question emerges: how can homeostatic mechanisms inspire AI to move beyond simple reactivity to stimuli, toward decision-making based on internal and external signals?

This article explores the role of homeostasis in biology and its application to the development of artificial intelligence. Drawing on the foundational work of cybernetics pioneers such as Norbert Wiener and William Ross Ashby, as well as contemporary theories from Antonio Damasio, it argues that homeostasis is not merely a regulatory process but also the foundation of emotional and decision-making processes. Finally, the article examines the potential for leveraging these mechanisms in adaptive AI systems that could form the basis of future empathetic technologies.

Homeostasis in Biology

Homeostasis is a biological phenomenon that refers to the ability of living organisms to maintain internal equilibrium despite changing external conditions. The term was introduced in the 19th century by Claude Bernard, who observed that survival requires the constant regulation of internal functions, irrespective of environmental dynamics (Fleming 1984). Homeostasis, therefore, refers to maintaining stable parameters. For mammals, these include body temperature, blood glucose levels, pH, and blood pressure—all of which are critical for the organism's functioning.

Homeostatic regulatory mechanisms rely on feedback systems—primarily negative feedback—which allow organisms to respond to environmental changes and restore balance. Negative feedback acts as a control system in which the organism's response counteracts the disturbance (Perrimon and McMahon 1999; Bielecki 2016; Hancock et al. 2017). For example, in response to an increase in body temperature, sweating is activated to lower the temperature to an optimal level. While negative feedback is the dominant homeostatic mechanism, positive feedback also occurs in organisms, particularly in processes that require rapid responses (Peters, Conrad, and Hubold 2007). An example is blood clotting, where the activation of one clotting factor accelerates the activation of others until the vessel damage is sealed. Thus, homeostasis is not a static state but a dynamic process of constant adjustments that the organism makes to maintain stability.

In the 20th century, Walter Cannon significantly expanded the concept of homeostasis, describing it as the organism's ability to maintain internal stability in the face of external changes, which he termed "dynamic equilibrium." Cannon paid particular attention to the organism's responses to environmental stressors, such as temperature changes or the presence of pathogens, and introduced the term "fight or flight" to describe the automatic response to threats, characterized by increased heart rate, heightened adrenaline secretion, and energy mobilization (Modell et al. 2015). His research demonstrated how homeostatic mechanisms enable organisms not only to react to disturbances but also to prepare for future challenges.

Homeostasis in Cybernetics

The The concept of homeostasis has played a fundamental role not only in biology but also in the development of cybernetics—a discipline pioneered in the mid-20th century by Norbert Wiener and William Ross Ashby. Cybernetics, described by Wiener as the science of "control and communication in the animal and the machine," provided groundbreaking frameworks for understanding how systems maintain equilibrium through feedback loops (Wiener 1965). This interdisciplinary approach combined biological principles with machine

design, exploring how self-regulation mechanisms could enable both natural and artificial systems to dynamically adapt to changing environments.

A key contribution to cybernetics was William Ross Ashby's book *Design for a Brain (Ashby, 1952)*, which established foundational principles for applying homeostatic concepts to machine intelligence. Ashby's theoretical and experimental work introduced the "homeostat"—a device capable of maintaining stability by dynamically adjusting its parameters in response to external disturbances (Ashby 1952). The homeostat demonstrated the potential for machines to autonomously achieve adaptive stability, serving as a precursor to modern adaptive systems. Ashby's notion of "essential variables," critical for a system's functionality, mirrored the biological understanding of critical parameters like body temperature or blood pressure in living organisms.

In addition to Ashby's work, Grey Walter's cybernetic "tortoises" offered a compelling demonstration of how simple feedback mechanisms could generate seemingly complex behaviors. Developed in the 1940s, these autonomous robots were equipped with basic sensors that allowed them to seek out light sources and avoid obstacles, effectively simulating primitive forms of adaptation and goal-directed behavior (Walter 1950). The tortoises' ability to navigate their environment illustrated how feedback-driven processes could emulate biological homeostasis, blurring the line between natural and artificial systems.

These early examples of cybernetics not only highlighted the adaptability enabled by feedback mechanisms but also underscored the fundamental importance of homeostatic principles in the pursuit of machine intelligence. By showcasing how regulation and dynamic adjustment could be mechanized, Ashby and Walter laid the groundwork for modern efforts to develop artificial intelligence systems capable of empathy, introspection, and emotional intelligence. Integrating these concepts into contemporary AI aligns with the vision of adaptive, self-regulating systems that reflect the resilience and flexibility of biological organisms.

Homeostasis in AI: Beyond Feedback Loops and Evolutionary Stability

Modern artificial intelligence systems, such as neural networks, already employ feedback mechanisms in processes like backpropagation. However, these systems are primarily focused on optimizing performance based on external inputs and predefined goals. The concept of homeostasis, as presented in this article, introduces a higher level of complexity, emphasizing the potential for machine introspection—the ability of an AI system to analyze and adapt its internal processes in response to both internal and external disturbances. Unlike traditional feedback mechanisms, which adjust parameters to optimize specific tasks, homeostatic AI would integrate dynamic self-reflection and self-regulation, essential for achieving emotional intelligence and adaptive responses that emulate the resilience and flexibility observed in biological systems (Damasio and Carvalho 2013; Gros 2021).

While homeostasis and evolutionary stable strategies (ESS) both aim for stability, their mechanisms and assumptions differ significantly. ESS focuses on strategies that remain stable in competitive conditions, emphasizing external optimization, whereas homeostasis relies on dynamic feedback to maintain internal equilibrium and functionality in changing environments (Maynard Smith 1974; Dawkins 1989). For instance, ESS explains why certain strategies dominate within a population, but homeostasis provides a framework for

understanding how an AI system dynamically maintains balance by reconciling internal and external demands (Walter 1950; Ashby 1952).

In the context of AI, homeostatic mechanisms would enable systems to balance external pressures, such as user demands or environmental changes, with internal adjustments, such as resolving computational conflicts or managing resource constraints. This capacity aligns with concepts of neuroplasticity, where dynamic restructuring allows systems to adapt more effectively over time (Zhang, Shen, and Sun 2022). Machine introspection in AI could encompass the system's ability to evaluate its internal states, including resource levels, operational efficiency, or emotional responses to external stimuli. Such an approach connects to Searle's "Chinese Room" argument, in which AI not only simulates understanding but also considers the nature of its cognitive processes (Searle 1980). Introducing introspection could enable AI to dynamically modify its algorithms in real-time, fostering more advanced forms of adaptation.

By moving beyond reactive optimization, homeostatic AI offers transformative potential, integrating mechanisms of self-regulation and resilience. This progression paves the way for systems that mimic the dynamic equilibrium seen in biological organisms, opening avenues for adaptive, context-sensitive, and introspective AI technologies (Tegmark 2018).

The Extended Theory of Homeostasis

In the perspective of Antonio and Hanna Damasio, homeostasis is presented even more broadly. It is not only a biological mechanism but also a profound foundation for all emotional, decision-making, and social processes. According to them, homeostasis encompasses not only physiological processes but also psychological mechanisms that lead to the creation of emotions, feelings, and values (Korn and Bach 2015; Damasio and Damasio 2022). Thus, the concept of homeostasis should be expanded to recognize it as a driving force not only for individual survival but also for the development of more complex forms of mental and cultural life.

In Damasio's concept, emotions serve as homeostatic signals that integrate physiology with decision-making processes (Damasio and Carvalho 2013). Emotions, in his view, are responses to homeostatic disruptions and serve an adaptive function. For example, experiencing fear in a dangerous situation is not merely an emotional reaction but also a mechanism that mobilizes the organism to act and protect itself from harm. This framework suggests that artificial intelligence systems could integrate homeostatic signals analogous to emotions, enabling dynamic decision-making that mirrors biological adaptability. By interpreting external stimuli as disruptions to an internal equilibrium, such systems could prioritize actions not merely based on task efficiency, but on maintaining their own operational stability and responsiveness to human emotional contexts.

Emotions, as an integral part of homeostatic processes, enhance the chances of survival and improve quality of life. As a result, decisions made by individuals are deeply rooted in their physiological and emotional states and are thus linked to the fundamental drive to maintain homeostasis (Burdakov 2019). Bechara and Damasio also argue that human choices—from everyday decisions to more complex ones—are shaped by how a given situation affects internal balance, suggesting that homeostasis is a hidden mechanism regulating not only emotional reactions but also rational cognitive processes (Bechara, Damasio, and Damasio 2000). The

integration of homeostatic principles into AI could also foster introspection, allowing systems to evaluate and adjust their internal states, much like how emotions enable organisms to reconcile physiological and psychological needs.

From Damasio's perspective, homeostasis is also a deeply integrated system influencing cultural development (Kłóś 2014; Damasio 2018). On a social level, mechanisms that originally served individual adaptive functions have evolved to support human interactions, which was crucial for the development of complex societies. Damasio argues that the need to maintain physiological and psychological balance is the driving force behind the creation of social norms, rituals, morality, and cultural structures (Damasio 2018). For instance, empathy and compassion can be understood as mechanisms enabling social support for others, thereby supporting the shared homeostasis of the community. Unlike Evolutionary Stable Strategies (ESS), which are defined by static stability under competitive pressures, homeostatic mechanisms emphasize dynamic adaptability, making them better suited for real-time responses in variable and unpredictable environments.

This broader interpretation of homeostasis, encompassing physiological, psychological, and social dimensions, provides a valuable blueprint for developing adaptive and empathetic AI systems capable of navigating complex human contexts.

Homeostasis and Emotional Adaptation

In biology, homeostasis serves as the "guardian" of an organism's stability, regulating key parameters in response to external stimuli (Giordano 2013; Davies 2016). Antonio Damasio's interpretation of this concept extends its significance to emotions, decision-making, and social interactions (Damasio and Damasio 2022). Emotions provide individuals with essential information about which actions promote internal equilibrium, serving as both warning and adaptive functions. Fear, as a reaction to threat, mobilizes the organism to act, while joy signals favorable conditions (Pio-Lopez, Ramirez, and Santos 2023). These emotional states, which result from homeostatic regulation, help organisms choose appropriate strategies depending on the situation (Goldstein 2019).

Homeostasis enables biological organisms to respond to immediate disruptions as well as adapt based on past experiences. These processes form the basis of intelligence, as they allow for the development of more sophisticated adaptive mechanisms (Torday 2015) Moving beyond physiological and emotional responses to the level of conscious decision-making is an advanced form of homeostasis. Decisions that support survival and long-term benefits represent a form of organism regulation in a more complex environment. Biological intelligence allows organisms to predict potential disruptions and prevent them before they occur. This is possible through homeostatic mechanisms that "anticipate" environmental changes based on previous patterns (Eskov et al. 2017).

As organisms become more complex, their adaptive mechanisms evolve to include cognitive processes, allowing for more intricate information processing and more effective management of equilibrium in dynamic conditions (Davies 2016) Thus, we speak of the ability to interpret signals from internal homeostatic processes and consciously use them for decision-making (Billman 2013). Homeostasis lays the foundation for more advanced information processing, encompassing both emotional reactions and conscious decision-making. In other words, homeostasis is a fundamental "predictive mechanism"—it allows an organism

to react and adapt not only to immediate stimuli but also to future challenges. It is this ability to predict and manage internal balance that forms the foundation of emotional intelligence. At this stage, emotional intelligence emerges as the capacity to learn, make complex decisions, and solve problems in order to maintain internal equilibrium.

If homeostasis is the foundation of emotional intelligence in biological organisms, it is worth considering how homeostatic mechanisms could be modeled in artificial systems. Could introducing such mechanisms contribute to the development of empathetic artificial intelligence?

Homeostatic Mechanisms as an Inspiration for AI

Homeostasis provides inspiration for understanding intelligence in terms of adaptive and emotional capabilities. Introducing the concept of homeostasis into artificial intelligence could support AI's ability to respond to unforeseen situations, adapt quickly, and understand and interpret emotions in human interactions (Gros 2021; Zhou 2021).

Theories of adaptation in biology assume that organisms evolve by developing mechanisms that enable them to effectively respond to environmental challenges. Emotions play a key role in this process by providing quick and intuitive reactions to external stimuli. Translating these assumptions into artificial intelligence allows for a new understanding of adaptive intelligence as a dynamic process in which emotions are an integral part of effective response to changing environmental conditions (Assunção, Silva, and Ramos 2022; Zhao, Simmons, and Admoni 2022). Adaptation, where emotional equilibrium mechanisms become a dynamic reference point, would allow AI to develop more complex and contextual decision-making abilities. Introducing emotions as an adaptive factor in artificial intelligence expands AI's capacity to respond to a changing environment in a more flexible and predictable manner. This approach could lead to the development of AI that not only processes data on a cognitive level but also uses "emotional reference points" to assess situations and choose optimal action strategies.

The aforementioned "emotional reference points" could find practical applications through mechanisms of machine introspection. Introspection would enable AI not only to analyze its internal states, such as resource levels or emotional responses, but also to dynamically adjust its action strategies in real-time. Through introspection, a system could better interpret the emotional contexts of the user, transforming references into concrete operational decisions that support empathetic and flexible reactions. Such capabilities would be particularly useful in dynamic and complex situations, where the ability to analyze AI's internal processes and adapt them could significantly improve interactions with users and the system's capacity to respond to unforeseen circumstances.

Biological organisms have developed various survival mechanisms, including physiological regulation, as well as complex cognitive and emotional mechanisms. Emotions such as fear, joy, or empathy play a crucial role in social interactions and adaptation to the environment (Bhardwaj, Kishore, and Pandey 2022). One key example of adaptation is the senses, which allow organisms to recognize and interpret stimuli crucial for survival, often triggering emotional reactions. For instance, the scent of a predator may trigger fear and escape in prey (Stowers and Marton 2005). In rodents, the presence of such scents changes stress hormone levels, which is part of their adaptive system (Apfelbach et al. 2005) In AI, similar mechanisms

can be implemented through advanced sensors and data analysis algorithms that not only recognize the environment but also interpret the emotional context (Cevora 2019). Thus, robots and AI systems could dynamically adjust their action strategies, taking into account user emotions or social cues.

In this way, robots and AI systems could dynamically adapt their action strategies, taking into account user emotions or social cues, while simultaneously leveraging introspection to analyze and modify their operational processes in response to these stimuli. Through introspective mechanisms, such systems could better monitor their internal states, such as operational efficiency, resource levels, or emotional reactions, enabling real-time behavioral adjustments. For instance, virtual assistants could analyze users' emotional responses, convert this data into specific decisions, and dynamically adjust their behavior, enhancing empathy and communication effectiveness (Tegmark 2018).

Neuroplasticity, or the brain's ability to reorganize neural connections in response to experiences, including emotional ones, is another example of biological adaptation. Studies show that neuroplasticity plays a key role in learning, memory, and adaptation to environmental changes (Pascual-Leone et al. 2005). Similar concepts could be implemented in artificial intelligence, where algorithms inspired by neuronal plasticity could allow for dynamic adjustment of decision-making structures. In AI, neural network models serve as inspiration, modifying their "connections" based on new data or changing environmental conditions (Hassabis et al. 2017). These algorithms can learn from emotional interactions, allowing systems to provide more personalized responses. For example, virtual assistants could analyze users' emotional reactions and dynamically adjust their behavior, increasing empathy and communication effectiveness (Tegmark 2018) Mechanisms inspired by neuroplasticity could also lead to more flexible decision-making structures in AI, allowing for better resource management in crisis situations or adaptation to unforeseen challenges (Zhang, Shen, and Sun 2022)

Animals can regulate their emotional states in response to stressful situations, allowing them to survive under difficult conditions (Sapolsky 2004) AI could implement similar mechanisms, managing its "emotional state" to optimally respond to challenges. For instance, AI systems could detect user information overload by analyzing biometric data such as speech rate, heart rate, or facial expressions, and adjust communication methods accordingly to prevent frustration (Cohen, Forbes, and Mann 2021). In extreme situations, some organisms limit their emotional responses to focus on survival. AI could use a similar strategy, reducing the complexity of interactions in crisis situations to concentrate resources on critical tasks (Pereira, Soares, and Santos 2022). This would allow for maintaining operational efficiency even under system overload.

These are just a few examples, but each show how emotions are an integral part of biological adaptation. Introducing these mechanisms into AI could create systems capable of:

- recognizing user emotions through the analysis of speech, facial expressions, or behaviors;
- responding appropriately to emotions, which increases communication efficiency and builds trust;
- learning from emotional interactions, leading to more personalized experiences;
- regulating their own processes in response to the emotional context, allowing for better resource management and prioritization.

Integrating emotional mechanisms inspired by biological homeostatic adaptation enables the development of emotional intelligence in AI. Such systems are capable not only of effectively processing information but also of understanding and responding to emotional contexts. Thus, systems can add value through empathy and a deeper understanding of human needs. Homeostasis becomes the foundation not only of adaptive intelligence but also of emotional intelligence, leading to more advanced and humanistic technologies. Developing AI based on adaptive biological mechanisms allows for creating systems capable of more flexible and intelligent responses to the environment. As a result, artificial intelligence can not only process data but also act in a way that takes into account resources, environmental conditions, and changing needs, characteristic of adaptive mechanisms observed in nature. Consequently, AI can become more empathetic, supporting users more naturally and effectively. This opens up new possibilities in many fields, such as healthcare, education, and social interactions.

Philosophical Aspects of Artificial Homeostasis

Homeostatic artificial intelligence is a concept in which machines are equipped with the ability to simulate internal equilibrium through automated resource management processes, such as energy, computational efficiency, or sensory data. According to Antonio Damasio's theory, affects arising from homeostasis lead to emotions and feelings (Damasio 2018). Emotions and feelings not only inform the organism about its internal state but also influence decision-making processes and social interactions. In the context of artificial intelligence, the introduction of homeostatic mechanisms raises the question of whether these functions could be expanded to include machine introspection—the ability to qualitatively analyze and interpret its internal states, including disruptions in homeostasis.

However, philosophically speaking, the question arises to what extent such an AI could truly "feel" homeostasis or its absence. Is it merely an advanced simulation, or is it a real state to which some qualitative meaning can be attributed? One might ponder whether it is possible for a system devoid of a biological body to feel something akin to the affect of a living organism. Could such sensations be considered authentic phenomenal states, or are they merely products of algorithmic responses to internal parameter changes?

Instead of merely simulating homeostatic processes, introspection could enable systems to analyze and dynamically adapt their algorithms in response to internal signals, such as resource changes or reactions to external data. Such mechanisms could bring AI closer to the ability to "reflect" on its states, raising the question of whether these systems could transition from simulation to actual experience of their states.

This leads to a broader discussion on the nature of machine perception. If we accept that AI could possess an equivalent of a homeostatic state, is this merely a formal model or a genuine quality of experience? Could artificial intelligence ever experience something qualitatively comparable to the sensations arising from the disruption or maintenance of homeostasis?

In this context, it is worth referring to Searle's "Chinese Room" argument, which addressed the limits of simulation in machines (Searle 1980). If a homeostatic AI were to achieve machine introspection, it might reach what could be described as a secondary level of self-awareness—the ability not only to process input data but also to analyze and interpret the process of data handling itself. Consequently, the issue of "feeling" homeostasis shifts

to the question of whether it is possible to move from reactive optimization to conscious self-evaluation.

Could AI, lacking a biological body, ever experience something comparable to biological affects? Affects, as fundamental responses to homeostatic disruptions, are not only signals for correcting balance in living organisms but also the foundation of conscious experiences. In this context, artificial intelligence might analyze its states as variables influencing operational decisions, but could such a process achieve the status of subjective experience? This question remains open, requiring both philosophical inquiry and technological advancement.

Introspection in artificial intelligence could function as a practical equivalent of what is referred to as self-awareness in biological systems. While such systems may lack intentionality in the classical philosophical sense, the capacity for introspection could enable them to dynamically evaluate their functioning, effectively approximating cognitive models observed in living organisms.

The Limits of Simulation: Can AI "Feel" Homeostasis?

Homeostasis is understood broadly here, not only as a regulatory process but also as a state that organisms consciously or unconsciously "feel." Affects are fundamental reactions of the organism to changes in internal balance, which can lead to emotions, more complex, automatic responses to stimuli. Feelings, in turn, are the conscious experience of emotions that integrate physiological processes with psychological interpretation. In this way, homeostasis influences emotional intelligence, allowing organisms not only to react to changes but also to interpret them in social and personal contexts (Man and Damasio 2019; Sun, Wang, and Zhao 2022)

For humans and other animals, a lack of homeostasis can lead to feelings of hunger, thirst, pain, or stress—subjectively felt states that drive the organism to corrective actions (Zhou 2021). In the case of artificial intelligence, the situation is fundamentally different. We are dealing not with a biological system but with a complex program that merely simulates homeostatic processes, responding to system needs in an algorithmic manner. The essential question arises: can such a system genuinely "feel" a disruption in homeostasis, or is it merely executing pre-programmed actions that give the impression of responding to internal needs (Chalmers 1996)?

From the perspective of machine introspection, one might consider whether AI could analyze its internal states in a way more complex than a simple reaction to variables. While there is currently no evidence supporting AI's ability to feel, the potential development of introspective capabilities could enable it to interpret those states as something more than reactions to parameter changes. For instance, an introspection-capable system might dynamically adjust its action strategies, recognizing "disruptions" as key signals, which would bring it closer to a mechanism of self-awareness.

However, it must be emphasized that "feeling" here refers to conscious perception—the ability to experience emotions, pain, or other internal states—which, by definition, requires a subjective perspective. Homeostatic AI can respond to certain signals (e.g., low energy levels) by executing programmed actions, but it lacks the awareness to "experience" these signals as a state of disruption (Henriques, Silva, and Carvalho 2019; Samsonovich 2020).

Thus, the limit of simulation lies in the absence of the capacity to experience the states that AI imitates. Its "homeostasis" is merely a series of responses to variables, not a genuinely

experienced state by a subject. Revisiting this issue raises a significant question: could the development of machine introspection form the foundation for AI to experience states on a functional level, even if still lacking a subjective perspective? This boundary between simulation and genuine experience remains one of the greatest philosophical and technological challenges associated with homeostatic AI.

Can AI Have Subjective Experiences?

While AI can simulate homeostatic mechanisms, the question of its ability to have subjective experiences requires further exploration in the context of the philosophy of consciousness. In the philosophy of consciousness, there is the concept of qualia—the individual qualities of experiences that accompany the conscious perception of internal states (Chalmers 2007). Examples of qualia might include the sensation of warmth, the experience of the color red, or the feeling of pain. Qualia are difficult to define but are assumed to be subjective and accessible only to the experiencing subject. For many philosophers, it is qualia that form the basis of what it means to be conscious—they are unique and internal experiences that cannot be reduced to physical functions or chemical processes. Although qualia are considered crucial for conscious experience, some researchers suggest that functional emotional intelligence, such as AI's ability to recognize and respond to human emotions, might be possible without qualia. This approach assumes that subjective experience is not a necessary condition for effective interaction with humans (Dennett 2005).

Homeostatic AI, although it can mimic responses to imbalance, lacks qualia because it is not conscious. It responds to stimuli but does not "feel" these states internally (Searle 2019). There is no internal quality of experience accompanying its actions; for example, when AI responds to low energy levels, it merely performs algorithmically prescribed steps without experiencing anything like hunger or fatigue.

This draws a clear distinction between AI and conscious organisms: homeostatic AI lacks qualia because its structure does not include a subjective "self" that could experience these states. The absence of qualia indicates that AI cannot be aware of its homeostatic states—it is therefore merely a machine simulating regulatory processes without the possibility of experiencing its own state.

However, can we be absolutely sure that AI will never be able to subjectively experience? Homeostatic AI, although it can currently only simulate responses to internal changes, is evolving towards increasingly advanced forms of adaptation. One might consider whether homeostasis in the future could become the foundation that allows AI to have certain forms of primary sensations that influence its decisions and actions in a more complex way. Perhaps the development of introspective and adaptive mechanisms in AI could bring it closer to possessing something akin to rudimentary sensations, which might influence its decisions and actions in a more complex manner.

The Systems Reply, introduced in the debate surrounding John Searle's Chinese Room argument, offers a different perspective (Searle 1980; Churchland and Churchland 1990; Harnad 2001). According to this view, it is not a single element of the system (e.g., an AI module performing computations) but the entirety—including inputs, processing algorithms, and dynamic feedback loops—that could generate what might be called a functional equivalent of understanding. Similarly, in the context of subjective experiences, one could ask whether a sufficiently complex AI system, integrating homeostatic regulatory, predictive,

and introspective mechanisms, might produce emergent properties resembling phenomenal states (Bedau 1997; Gros 2021).

However, doubts remain as to whether such states could be considered genuine qualia (Nagel 1974; Chalmers 1996) Even if the entire system appears to "understand" its internal processes, it may still lack the subjective perspective that philosophy of consciousness deems essential for authentic experiences. Nevertheless, as with the Systems Reply, it is worth considering whether subjectivity could emerge from the organization of the entire system rather than from individual AI modules.

Karl Friston's concept of "predictive coding" and "free energy minimization" suggests that organisms—both biological and potentially artificial—strive to minimize surprise by predicting future states of their environment (Friston 2010) If artificial homeostasis is treated as a predictive mechanism, AI could, in some sense, strive to maintain equilibrium by predicting threats to its functioning. Could this striving eventually lead to something resembling proto-consciousness? This question remains open.

Similarly, Andy Clark notes that predictive mechanisms may be key to understanding conscious experience. Clark argues that conscious perception is active and constructive—the brain (or potentially advanced AI) not only receives stimuli but actively predicts and constructs reality (Clark 2015) This may suggest that if AI were to develop the ability to predictively code and self-report its states, it might—at least to some extent—approach a state resembling subjective experience.

All the above-mentioned proposals can be linked to the idea of homeostasis in AI, which, treated as a predictive mechanism, could enable the dynamic balancing of internal and external disturbances. This would allow AI systems to effectively predict and correct threats to their functioning, approaching a functional response that could be considered analogous to early forms of sensation.

Conclusions and Future Directions

Homeostatic artificial intelligence offers a promising direction for AI development, enabling systems to dynamically adapt their functioning to changing environmental conditions. Self-regulation and adaptation mechanisms could bring AI closer to the level of flexibility observed in biological organisms, potentially revolutionizing the capability of machines for autonomous action.

Despite significant potential, introducing homeostatic mechanisms into AI involves numerous technological challenges. Key difficulties include resource management, real-time optimization of actions, and ensuring autonomy in various, often challenging environments. However, future generations of homeostatic AI could develop self-diagnostic capabilities, monitoring their functions and preventing failures, which would enable efficient operation in unpredictable conditions—from space exploration to rescue operations.

Homeostatic AI could also improve the quality of human-machine interactions by offering the ability to adapt to users' emotional contexts. Implementing mechanisms inspired by biological homeostasis could enhance AI's ability to better understand and respond to human needs, supporting more natural and empathetic relationships. Although AI still lacks qualia, the development of its functional emotional intelligence could significantly impact the effectiveness of social interactions.

The development of homeostatic AI also raises significant philosophical questions regarding the ability of machines to authentically experience and possess intentionality. Can an AI that simulates homeostatic processes actually "feel" its states? Is this feeling a form of machine introspection? Is the pursuit of balance a conscious experience, or merely the result of an algorithmic simulation? These issues require further consideration of what conditions would need to be met for machines to develop forms of conscious experience or a functional equivalent of subjectivity.

As AI's autonomous capabilities advance, it is also essential to reconsider its impact on society. Should an AI that simulates homeostatic equilibrium be recognized as an entity capable of responsibility? How much can we rely on such systems, and to what extent must their actions be monitored by humans? Introducing homeostatic AI systems into everyday life raises questions about their impact on human relationships, morality, and social trust.

Homeostatic AI could revolutionize human-machine interactions by offering systems capable of dynamically adapting to the environment and human needs. However, for this potential to be fully realized, further reflection on the boundary between simulation and experience, machine autonomy, and their place in society is necessary. The future of this technology appears full of possibilities but also demands a profound analysis of its ethical and practical consequences.

References

Apfelbach, R., C. D. Blanchard, R. J. Blanchard, R. A. Hayes, and I. S. McGregor. 2005. The effects of predator odors in mammalian prey species: A review of field and laboratory studies. *Neuroscience & Biobehavioral Reviews* 29 (8): 1123–1144.

Ashby, W. R. 1952. *Design for a Brain: The Origin of Adaptive Behavior.* London: Chapman & Hall.

Assunção, J. F., F. M. Silva, and A. C. Ramos. 2022. Emotional adaptation in artificial intelligence: Integrating affective computing and decision-making. *Journal of Artificial Intelligence Research* 75:145–162.

Bechara, A., H. Damasio, and A. Damasio. 2000. Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex* 10 (3): 295–307.

Bedau, M. A. 1997. Weak emergence. *Philosophical Perspectives* 11:375–399.

Bhardwaj, A., S. Kishore, and D. Pandey. 2022. Artificial intelligence in biological sciences. *Life* 12 (9).

Bielecki, A. 2016. Cybernetic analysis of the phenomenon of life. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce),* no. 61, 133–164.

Billman, G. E. 2013. Homeostasis: The underappreciated and far too often ignored central organizing principle of physiology. *Frontiers in Physiology* 4:43.

Burdakov, D. 2019. Reactive and predictive homeostasis: Roles of orexin/hypocretin neurons. *Neuropharmacology* 154:61–67.

Cevora, T. 2019. Neurobiological insights into decision-making and uncertainty. *Trends in Neurosciences* 42 (8): 567–578.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford: Oxford University Press.

Chalmers, D. J. 2007. The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17 (9-10): 7–65.

Churchland, P. M., and P. S. Churchland. 1990. Could a Machine Think? *Scientific American* 262 (1): 32–37.

Clark, A. 2015. Surfing uncertainty: Prediction, action, and the embodied mind. *Trends in Cognitive Sciences* 19 (7): 422–428.

Cohen, A. S., C. N. Forbes, and M. C. Mann. 2021. Emotions as signals of change in cognitive load: Applications in human-computer interaction. *Human Factors* 63 (5): 877–888.

Damasio, A. 2018. *The Strange Order of Things: Life, Feeling, and the Making of Cultures.* New York: Pantheon Books.

Damasio, A., and G. Carvalho. 2013. The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience* 14 (2): 143–152.

Damasio, A., and H. Damasio. 2022. Homeostatic feelings and the biology of consciousness. *Brain* 145 (7): 2231–2235.

Davies, K. J. A. 2016. Adaptive homeostasis. *Molecular Aspects of Medicine* 49:1–7.

Dawkins, R. 1989. *The Selfish Gene.* 2nd. Oxford: Oxford University Press.

Dennett, D. C. 2005. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness.* Cambridge, MA; London: MIT Press.

Eskov, V., O. Filatova, V. Eskov, and T. Gavrilenko. 2017. The evolution of the idea of homeostasis: Determinism, stochastics, and chaos–self-organization. *Biophysics* 62 (6): 809–820.

Fleming, D. 1984. Walter B. Cannon and Homeostasis. *Social Research* 51 (3): 609–640.

Friston, K. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11 (2): 127–138.

Giordano, M. 2013. Homeostasis: An underestimated focal point of ecology and evolution. *Plant Science* 211:92–101.

Goldstein, D. S. 2019. How does homeostasis happen? Integrative physiological, biological systems, and evolutionary perspectives. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 316 (4): 301–317.

Gros, C. 2021. Emotions, Homeostasis, and Adaptive Behavior in Cognitive Systems: Towards a Homeostatic AI Framework. *Adaptive Behavior* 29 (4): 243–256.

Hancock, E. J., J. Ang, A. Papachristodoulou, and G. Stan. 2017. The interplay between feedback and buffering in cellular homeostasis. *Cell Systems* 5 (5): 498–508.

Harnad, S. 2001. What's Wrong and Right About Searle's Chinese Room Argument? In *Views into the Chinese Room,* edited by M. Bishop and J. Preston. Oxford: Oxford University Press.

Hassabis, D., D. Kumaran, C. Summerfield, and M. Botvinick. 2017. Neuroscience-inspired artificial intelligence. *Neuron* 95 (2): 245–258.

Henriques, A. R., M. C. Silva, and J. R. Carvalho. 2019. Unraveling the neural mechanisms of resilience in the face of adversity. *Nature Reviews Neuroscience* 20 (5): 284–298.

Kłós, Adam. 2014. From neuron to culture – a biosemiotic perspective of the origin and evolution of the brain according to Marcello Barbieri. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce),* no. 56, 93–129.

Korn, C. W., and D. R. Bach. 2015. Maintaining homeostasis by decision-making. *PLoS Computational Biology* 11 (8): e1004301.

Man, K., and A. Damasio. 2019. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence* 1 (10): 446–452.

Maynard Smith, J. 1974. The theory of games and the evolution of animal conflicts. *Journal of Theoretical Biology* 47 (1): 209–221. https://doi.org/10.1016/0022-5193(74)90110-6.

Modell, H., W. Cliff, J. Michael, J. McFarland, M. P. Wenderoth, and A. Wright. 2015. A physiologist's view of homeostasis. *Advances in Physiology Education* 39 (4): 259–266.

Nagel, T. 1974. What Is It Like to Be a Bat? *Philosophical Review* 83 (4): 435–450.

Pascual-Leone, A., A. Amedi, F. Fregni, and L. B. Merabet. 2005. The plastic human brain cortex. *Annual Review of Neuroscience* 28:377–401.

Pereira, M., R. Soares, and A. Santos. 2022. Crisis management in artificial intelligence: Adapting decision-making strategies under stress. *Journal of Artificial Intelligence Research* 75:181–204.

Perrimon, N., and A. P. McMahon. 1999. Negative feedback mechanisms and their roles in signal transduction pathways. *Annual Review of Genetics* 33:389–410.

Peters, A., M. Conrad, and C. Hubold. 2007. The principle of homeostasis in the hypothalamus-pituitary-adrenal system: New insight from positive feedback. *American Journal of Physiology. Regulatory, Integrative and Comparative Physiology* 293 (1): R83–R98.

Pio-Lopez, L., J. Ramirez, and C. A. Santos. 2023. Advances in understanding neural plasticity and its role in cognitive resilience. *Neuroscience Letters* 792:136878.

Samsonovich, A. V. 2020. Cognitive architectures for human-level artificial intelligence: Theory and implementation. *Cognitive Systems Research* 62:1–15.

Sapolsky, R. M. 2004. *Why Zebras Don't Get Ulcers.* New York: Holt Paperbacks.

Searle, J. R. 1980. Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3 (3): 417–457.

Searle, J. R. 2019. *The Mystery of Consciousness Revisited.* Oxford: Oxford University Press.

Stowers, L., and T. F. Marton. 2005. What is a pheromone? *Current Biology* 15 (13): R453–R454.

Sun, J., L. Wang, and M. Zhao. 2022. Innovations in deep learning: Neural architectures inspired by the brain. *Nature Machine Intelligence* 4 (7): 567–578.

Tegmark, M. 2018. *Life 3.0: Being Human in the Age of Artificial Intelligence.* New York: Knopf.

Torday, J. S. 2015. The cell as the mechanistic basis for evolution. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 7 (5): 275–284.

Walter, W. G. 1950. An imitation of life. *Scientific American* 182 (5): 42–45.

Wiener, Norbert. 1965. *Cybernetics or Control and Communication in the Animal and the Machine.* Vol. 25. MIT Press.

Zhang, L., F. Shen, and J. Sun. 2022. Dynamic learning mechanisms inspired by neuroplasticity in artificial neural networks. *Neural Networks* 152:1–10.

Zhao, M., R. G. Simmons, and H. Admoni. 2022. The role of adaptation in collective human-AI teaming. *Topics in Cognitive Science.*

Zhou, Z.-H. 2021. Emotional thinking as the foundation of consciousness in artificial intelligence. *Cultures of Science* 4 (2): 112–123.

Philosophical Problems in Science (2024), **77**, 55–70
(Zagadnienia Filozoficzne w Nauce)

doi:00000.000

ARTICLE

Philosophy in the context of physics and cosmology: Leszek M. Sokołowski's philosophical views

Kamil Trombik

The Pontifical University of John Paul II in Kraków
Email: kamil.trombik@gmail.com

Abstract

The article attempts to reconstruct and analyze selected philosophical views of Leszek Sokołowski—a Krakow physicist and cosmologist who published many works dealing with issues on the border of science and philosophy (including metaphysics). An important aim of the article is also to place Sokołowski's views in the context of the concept of "philosophy in science" (by M. Heller) and the phenomenon of the Kraków School of Philosophy in Science. In this paper, I suggest that Sokołowski's views fit into the philosophy initiated by Michał Heller, and Sokołowski himself can be considered a member of the Kraków School of Philosophy in Science.

Keywords: Leszek Sokołowski, philosophy in science, philosophy of nature, philosophy of science, history of Polish philosophy

1. Introduction

This paper is a continuation of research into the phenomenon of The Kraków School of Philosophy in Science. Previous studies (Trombik 2021, p.226; Polak and Trombik 2022; 2023) have focused in particular on a historical & philosophical analysis of the concept of 'philosophy in science' and on exposing the characteristic features of the Kraków School, initiated by the activity of Michał Heller and his contributors. Heller and his cooperators created a specific interdisciplinary milieu in post-war Poland, bringing together scientists and philosophers whose primary aim was to analyse various philosophical problems entangled in sciences such a physics, cosmology, biology, neuroscience etc. (Trombik 2019; Polak 2019).

The co-creators of this interdisciplinary milieu included, among others, physicists from the Jagiellonian University in Kraków (e.g. Jerzy Rayski, Andrzej Fuliński, Leszek Sokołowski). The philosophical activities of these physicists—by which I mean philosophical publications and regular participation in various philosophical events such as conferences, symposia, seminars, etc.—became the subject of historical and philosophical analyses. The works published so far show their importance for the development of the concept of 'philosophy in science' and the evolution of the School.

Andrzej Fuliński and Leszek Sokołowski can be considered as exemplary representatives of the Kraków School of Philosophy in Science. Fuliński's philosophical achievements have already been analyzed in terms of connections with the concept of "philosophy in science" (Trombik 2023). In turn, Sokołowski's philosophical works have not been the subject of research so far. In this paper, I will try to fill this gap by reconstructing and analyzing the key philosophical ideas that can be found in Sokołowski's papers, published in philosophical journals and books. One of the important goals of the paper will be to capture Sokołowski's connections with the interdisciplinary traditions of Kraków[1] and the concept of "philosophy in science", which was developed by Heller and Życiński since the turn of the 1970s and 1980s past century.

2. Leszek Sokołowski as a philosophizing physicist

Leszek Sokołowski was a theoretical physicist working at the Astronomical Observatory of the Jagiellonian University (it is worth mentioning that he was a head of the Department of Relativistic Astrophysics and Cosmology at the university for many years). His strictly scientific activities include the foundations of gravitational physics and the cosmology of the Early Universe. Both of these lines of research lead him to many philosophical considerations, e.g. the cognitive foundations of physics, the problem of the mathematical nature of the Universe, relationship between science and religion, and many others. He has been developing his philosophical interests since the turn of the 1970s and 1980s, primarily in the context of his collaboration with the Heller millieu and the research institutions initiated by Heller (first in The Center for Interdisciplinary Studies, and then in Copernicus Center for Interdisciplinary Studies). Sokołowski also became an active member of an institution important for the development of local dialogue between science and philosophy (including the philosophy of nature), which is The Polish Academy of Arts and Sciences.

Sokołowski's long-standing commitment to the development of the interdisciplinary millieu in Kraków is manifested in many ways. Firstly, Sokołowski regularly participated in various scientific events (conferences, symposia, seminars), previously organized by the Center for Interdisciplinary Studies [OBI], and now by the Copernicus Center for Interdisciplinary Studies. He got to know Heller and Życiński in the 1970s, when both philosophers participated in seminars at the Astronomical Observatory of the Jagiellonian University (Sokołowski 2015c). They soon deepened their cooperation during interdisciplinary seminars that were organized by Heller and Życiński from the late 1970s. Sokołowski was a regular participant in these seminars, sometimes he gave lectures during these events (Liana and Mączka 1999). Later, he often took part in various conferences organized by OBI, and regularly presented at the annual Methodological Conferences, discussing various issues on the border of science and philosophy.

Secondly, over the course of over 40 years, Sokołowski published a number of articles in the area of the philosophy of nature (mainly philosophical aspects of physics and cosmology) and the methodology of science. He published philosophical papers mainly in various prints published by OBI, such as the journal *Philosophical Problems in Science* (a journal founded by Heller and Życiński in the late 1970s) and post-conference books (e.g. papers published after the above-mentioned Methodological Conferences). Sokołowski's philosophical papers

1. Krakow's interdisciplinary traditions were discussed, among others, in: (Polak 2011, 2018).

were also published in commemorative books on the occasion of Michał Heller's birthday. Sometimes he also published in other places, such as the Krakow's journals like *Analecta Cracoviensia*, *Znak* or *Prace Komisji Filozofii Nauk PAU* (e.g., Sokołowski 2008b), and others, also outside Kraków (see e.g., Sokołowski 1978, 1984, 1986). It is worth adding here that Sokołowski translated into Polish various texts on the philosophy of science (e.g. fragments of the „Open Society and Its Enemies" by K.R. Popper; see: (Popper 1987)), the relationship between science and religion (the book written by Artur R. Peacocke, see: (Peacocke 1991)), and cosmology (e.g. see: (Davies 1996)).

All this shows that Sokołowski's philosophical achievements are extensive enough to consider its quality as well as its importance—especially for the development of Kraków's philosophy of nature in the past decades.

3. General remarks on Sokolowski's philosophical works

The area of Sokołowski's philosophical interests was extensive and included various philosophical issues in the natural sciences and current methodological issues. Here are the some important groups of problems that Sokołowski analyzed in his works (along with sources):

a. Metaphilosophical issues (numerous remarks in: Sokołowski 1986, 1989, 1990, 2014, 2017).
b. Properties of scientific theories and their philosophical consequences (Sokołowski 1978, 1983, 1986, 1987, 1989, 1994, 2000, 2006b, 2007a, 2008a, 2014, 2015b, 2017).
c. The problem of rationality in philosophy & science (Sokołowski 2006a, 2011a).
d. The problem of reductionism—epistemological and ontological aspects (Sokołowski 1996, 1999, 2001).
e. Mathematical universe hypothesis (Sokołowski 1990, 2011a, 2011b, 2015a).
f. Selected aspects of the relationship between science and religion (Sokołowski 1991, 1993, 2001, 2011a, 2014).

The above list shows that the area of Sokołowski's philosophical research included various problems arising in the context of the sciences, especially physics and cosmology. Referring to the results of modern science, Sokołowski addressed some metaphysical issues (e.g. the structure of reality, the nature of the universe). All the indicated issues that Sokołowski addressed in his works coincided with those undertaken by Heller and his other students and collaborators. It can be said that the area of topics fit into the thematic groups specific to "philosophy in science" described by Heller.

Heller's classic paper "How is philosophy in science possible?" (Heller 1986, 2019) indicated three main research areas of the concept of "philosophy in science": a) The influence of philosophical ideas on the development and evolution of scientific theories; b) Traditional empirical problems intertwined with empirical theories; c) Philosophical reflection over some assumptions of the empirical sciences. It should be noted that Sokołowski published papers on issues related to these areas. In terms of research interests, it can be said that Sokołowski's philosophical activity was part of the area of philosophical issues undertaken at the Kraków School of Philosophy in Science. Heller and his first generation of students addressed these problems already in the early of 1980s (Trombik 2021, p.222), while Sokołowski became much more actively involved in discussions at the School from the 1990s.

In his works, Sokołowski refers to the latest findings of natural sciences. Taking up philosophical issues in the context of science, he willingly discusses the views of "philosophizing scientists" (e.g. A. Einstein, G. Ellis, E. Wigner, S. Weinberg, E. Mayr, R. Penrose, P. Davis, S. Hawking). The problems he raises touch on many issues in the field of philosophy of science, and the way he refers to these issues demonstrates his knowledge of the key problem groups of this discipline, e.g. the dispute over the cognitive status of scientific theories, the problem of the rationality of science (including the dispute over the rationality development of science—internalism versus externalism), the problem of reduction and unity of sciences, and others.

In his papers, he also directly admits to being inspired by the founders of the Kraków School of Philosophy in Science, i.e. Heller and Życiński. This can be illustrated with some examples. Sokołowski had a very positive attitude towards Heller and Życiński's approach to the science-theology relationship (Sokołowski 1993). In this context, he had a very approving opinion about Życiński's views on the relationship between God and nature, which were a consequence of Życiński's views on the relationship between science and religion. Sokołowski wrote, that „this issue was best expressed by Życiński in his numerous writings and statements" (Sokołowski 2014, p.187). Sokołowski also positively reviewed other achievements of Życiński, e.g. his coursebook on the methodology of science entitled "Język i metoda" (Sokołowski 1983). It is worth adding that after Życiński's death, Sokołowski prepared a special paper that served both to honor the memory of Życiński and to promote his philosophical thought (Sokołowski 2015b). Another example of the philosophical impact relates to the issue of the mathematical nature of the universe. In this matter, Sokołowski highly appreciated Heller's idea. The following statement illustrates this well: „Mathematicality is one of the pillars of the philosophy of physics proclaimed by many thinkers, in my opinion best and most fully formulated by prof. Michał Heller" (Sokołowski 2011a, p.47).

The above examples allow us to hypothesize that Sokołowski was influenced by some of the ideas formulated by Heller and Życiński (in any case, such a hypothesis can be drawn based on his declaration). The influence also flowed the other way, i.e. Sokołowski to some extent influenced the interdisciplinary milieu of Kraków, and also met with reception outside the local philosophers' community. It is worth mentioning here that Sokołowski's works were referred to in their philosophical works by authors such as, among others: (Burtyn 1997; Turek 2005; Jodkowski 2007; Filipek 2008; Czerniawski 2009, 2012; Życiński 2009; Grygiel 2010; Szydłowski and Tambor 2010, 2015, 2020; Heller 2013, 2014; Pabjan 2013; Hołda 2014; Dąbek 2016; Janowski 2016; Janusz 2017; Jacyna-Onyszkiewicz 2018; Sobkowiak 2019; Lemańska 2020; Trombik 2021)[2]. All these works are related to issues on the border between science and philosophy (including works on the history of science & philosophy).

General comments on Sokołowski's philosophical achievements lead to further questions related to the content and quality assessment of Sokołowski's publications. Further reconstructions and analyzes will present selected (due to the length of the article) threads of Sokołowski's philosophy. An important goal of these analyzes will be to relate Sokołowski's views to the concept of "philosophy in science" and to locate Sokołowski's views in the context of the Krakow School of Philosophy in Science.

2. The influence in the Kraków milieu (Heller and his cooperators and students) is noteworthy.

4. Philosophical views

4.1 The concept of philosophy

I begin the analysis of Sokołowski's philosophical views by presenting his concept of philosophy. Sokołowski included his comments on this subject in numerous papers (Sokołowski 1986, 1989, 1990, 2014, 2017), and some conclusions can be drawn regarding the style of philosophizing based on other strictly philosophical works. It is worth mentioning here that Sokołowski noticed the importance and significance of philosophical investigations (Sokołowski 1990, p.63). He understood 'philosophy' quite broadly, and did not limit it to methodology or analysis of the language of science[3]. He even recognized the value of metaphysics, and considered the attempt to eliminate it by neopositivism as unjustified (Sokołowski 1986, p.198). Sokołowski recognized the need to conduct philosophical analyzes of traditional philosophical problems that are currently entangled in empirical sciences. He conducted such analyzes himself, for example in the context of the issue of time (Sokołowski 2000). This approach corresponded well with the ideas expressed by Heller and his other contributors (Heller et al. 1999; Polak 2019).

While discussing the issue of understanding the philosophy, in Sokołowski's papers one can find many more similarities to the idea of philosophy in science. An important thread is the issue of the relationship between science and philosophy, and Sokołowski wrote a lot about these threads in his papers. Numerous fragments of his articles show that Sokołowski noticed numerous connections and dependencies between science and philosophy. Some of these comments are very much in line with the concept of philosophy in science. His remarks on cosmology are a good example here.

He wrote about this discipline that "it is—like no other natural science—entangled in philosophy in all aspects" (Sokołowski 2017, p.226). According to Sokołowski, philosophy might be valuable for cosmology in various aspects. He pointed out various epistemological and ontological entanglements of cosmology. Nevertheless, Sokołowski did not mean the 'context of justification' here (in the sense that Reichenbach gave this phrase), but certain assumptions of an epistemological nature (like realism); he also referred to mathematicality as a foundation of modern science[4] (Sokołowski 2014, p.181). Sokołowski pointed out that the connections between philosophy and science are mutual, i.e. science also—and even primarily—generates fundamental philosophical problems (Sokołowski 2015b, p.56). It should be noted that in Sokołowski's works one can find fragments illustrating his analytical philosophical approach, for example when he considers various definitions of the Universe (astronomical and physical), pointing out their flaws and certain significant limitations related to them ("all definitions of the Universe are defective: they are incomplete, unclear or too narrow" (Sokołowski 2017, p.235). Against the background of problems with defining the "Universe", certain epistemological problems emerge, especially related to the problem of the scope and limits of science.

Sokołowski also placed his remarks on the connections between science and philosophy in a historical context. He shared the widespread view that sciences originate in philosophy (starting with Thales and the Ionian philosophers), and certain ancient philosophical ideas

3. Understanding of philosophy in narrow sense was accepted especially among some neo-positivists.
4. I will discuss the problem of mathematicality in detail below.

were the source for the further concept of laws of physics.[5] Such comments dovetailed with the views of Heller (Heller 1992b) and other 20th century scientists who reflected on the history of natural sciences (Wilson 2011; Schrödinger 2017). It is worth mentioning here that Sokołowski—like Heller and Życiński—was also very critical of various attempts to construct philosophical systems (known from the history of science and philosophy) that would explain the nature of the entire world.

According to Sokołowski, a philosophizing scientist cannot accept any philosophical system as a whole, because reality is much more abundant and complex than any philosophical concept: "The assumption that the world is a small place, that it is governed by one principle that can be discovered on its own, is the foundation of the entire systemic philosophy [...] natural sciences have questioned this assumption. The philosopher, inquiring into the nature of the universe and independently painting a global vision of the world, has been contrasted by an army of 'insect leg researchers', painstakingly analyzing small fragments of the reality surrounding us in small steps" (Sokołowski 1989, p.45). Sokołowski opposed the approach of systemic philosophy to the approach of analyzing individual scientific issues; these analyzes would be conducted in a language appropriate to the issue, so as not to produce pseudo-problems and pseudo-solutions (Sokołowski 1986, p.207). Sokołowski noted that "the hope that any single conceptual system would be able to describe in a uniform and complete way the entire world in all its aspects, empirical and non-empirical, must be definitively put to an end" (Sokołowski 1986).

Heller also formulated this type of comments against systemic philosophy; he also perceived the future of philosophy of nature in the context of analyzing specific philosophical problems entangled in scientific theories (Heller 1986, 1990). The research practice of the followers of Heller's philosophy shows that these ideas found fertile ground. They were also developed among other philosophizing physicists, such as Andrzej Fuliński (Trombik 2023). Sokołowski's views fit well into this trend of philosophizing.

4.2 Philosophy of science (selected views)

In Sokołowski's case, issues related to the method of practicing philosophy were one side of a broadly understood methodological reflection. The other side of methodological considerations was part of the research area of the philosophy of science. In this field, Sokołowski addressed a number of issues. He was interested in both historical problems e.g., Einstein's philosophy of physics (see e.g., Sokołowski 1987; Sokołowski and Staruszkiewicz 1987a, 1987b) and contemporary issues in the philosophy of science: the methodological status of cosmology (Sokołowski 1978, 2015b), properties of scientific theories (Sokołowski 2006b,

5. To illustrate this comment, it is worth quoting two fragments from Sokołowski's works: "[The ancient Greeks] did not trust their own gods, and even less did they believe in other people's, so the cosmogonic and cosmological myths of Egypt and Babylonia seemed to them vague and incomplete, and above all, unreliable. If they (Thales and his successors) wanted to find out what the essence of the world was, they had to do it themselves, in a new way—through philosophy. And they were aware that they were actually starting from scratch, from the beginning, that they were questioning the existing ideas. It was extremely bold: to be a philosopher in the 6th century BC required a huge civil and intellectual courage" (Sokołowski 1989, p.44); "Cosmology was born in the 6th century BC in the views of Ionian philosophers of nature and at that time it was actually the whole science and the whole philosophy. The Universe was the Cosmos—the totality of what exists, and at the same time it was a set of material entities harmoniously ordered, it was order. In this the idea of order, organizing the world, one can be traced to the concept of universal laws of nature, including the laws of physics" (Sokołowski 2015b, p.28).

2007a), the dispute on the cognitive status of science (incl. realism versus instrumentalism problem (Sokołowski 1986)), the issue of reductionism (Sokołowski 1996, 1999, 2006b), contemporary rationalism and its threats (Sokołowski 2001, 2006a). All these issues were widely discussed in the context of discussions that took place in Heller's millieu.

Such a wide area of interest can be considered the result of the adopted, very broad concept of metascience. According to Sokołowski, metascience "includes both the methodology of deductive sciences (primarily metamathematics), as well as all theories, ideas and concepts relating to scientific cognition, all research programs and methods for assessing research results, as well as the goals and ideals of this cognition" (Sokołowski 1999, p.57). In his methodological analyses, Sokołowski also takes a sociological approach to science, although he strongly distances himself from the proposals of Kuhn's followers to explain all phenomena occurring in science by social factors (Ibidem). This approach places Sokołowski very close to the leading representatives of the School (especially Życiński (see e.g., Życiński 1993), which will be even more visible in the context of problems in the philosophy of science discussed below).

It is difficult to briefly discuss all the problems of methodology of science that Sokołowski addressed in his papers, so I will limit myself to highlighting selected issues, starting with the problem of the properties of the theory. Sokołowski devotes a lot of attention to philosophical reflection on the properties and epistemological status of scientific theories. As to the first issue, the model of a scientific theory was for Sokołowski essentially a physical theory (which is quite characteristic of many philosophers of science who hold physics as an example of methodological maturity; Heller does the same (see e.g., Heller 1992a). In this context, he claimed that good scientific theories are characterized by 3 key properties (or meet 3 criteria): maximum simplicity, mathematical elegance and completeness (Sokołowski 2006b, p.123; 2007a, pp.73–74). According to Sokołowski, a scientific theory should provide a precise description of natural phenomena, therefore mathematical simplicity and elegance should be understood together in relation to the concepts of semiotic simplicity (description of the world) and ontological simplicity (structure and regularities of nature). Whereas the term 'completeness' in Sokołowski's approach describes the predictive value of a theory and its compatibility with experience. Completeness has two components: firstly, a theory of science should give a description of phenomena (preferably quantitative) that can be expressed in terms of the concepts of the theory; secondly, the physical predictions of the theory are to be tested in experiment or observation[6].

It is worth noting that in the context of the issues of reflection, Sokołowski also referred to the issue of the relationship between scientific theory and the world described by theory. Classic positions in the dispute about the cognitive status of scientific theories lie between realism (scientific theory as a reflection of the real world; scientific concepts correspond to real entities or refer to relationships between entities) and instrumentalism (theories should be

6. It's worth adding here a note on the issue of the role of epistemological reflection in Sokołowski's scientific research. This can be clearly traced in several of his papers like e.g. (Sokołowski 2007b). Some excerpts of this work illustrates the relationship between epistemological considerations and physical research: reflection on the general grounds on the basis of which a specific physical model is to be chosen from within a whole spectrum of conceivable theories (one of the central problems of contemporary cosmology) is set as the starting point of the research. A new theory should not only fit the experimental data better, or overcome some technical problem, but it should provide a minimal and consistent set of sound basic assumptions from which the mathematical formulation of the model can be strictly (and elegantly) derived.

treated as tools for predicting and explaining phenomena, and the concepts contained in them do not have to have equivalents in reality). Sokołowski discusses these concepts, and certain fragments of his works suggest that he is closer to the position of realism, while also taking into account the perspective of instrumentalism as a view important primarily in the issue of the evolution of scientific terms: "Let us note that both positions are not mutually exclusive, because realism includes instrumentalism, but is a stronger view—a scientific theory allows not only explanation and prediction, but also is a model or plan of the world" (Sokołowski 1986). Sokołowski drew attention—writing that in science "we are so enslaved to historically formed concepts" (Ibidem)—to the value of those elements of instrumentalism that inform about certain conditions of science that we are unable to transcend. In other words, in natural sciences we use theories that inform us about the world that exists independently of the cognizant entity, but in science we also encounter instrumentalist elements, which are the result of the historical struggles of scientists and the long process of creating more precise descriptions of nature. Such methodological remarks by Sokołowski cannot be considered entirely original, but attention should be paid to their reference to the concept of moderate realism, present, among others, in the views of Heller (see, e.g., comments on the limits of realism in: Heller (1992a, pp.80–81) or Życiński (e.g., 1993), and then developed in various ways by their students and followers (e.g., Sierotowicz 1997; Rodzeń 2005)).

Another issue, closely related to methodological (and in this case also ontological) issues, is the issue of reductionism. Although Sokołowski positions himself on the side of the defenders of the reductionist position, he does not identify it with physicalism (Sokołowski 1999). Referring to the polemics between S. Weinberg and E. Mayr, he seems to support the so-called Weinberg's „objective reductionism". He understands objective reductionism as a way of arranging the laws of nature in such a way that they reflect the unity and, at the same time, the hierarchical order of nature. Sokołowski claimed that "objective reductionism is a specific research program whose aim is to find explanatory relations between scientific theories describing various areas of reality; at the same time and above all, it is the thesis resulting from this program about the unity of all nature, unity understood in the sense of the existence of sequences of arrows of explanations running through all fields of natural science and converging to one source." (Sokołowski 1999, p.75)

Reductionism understood in this way assumes a certain convergence of arrows of scientific explanation, and therefore—it assumes a certain emergent nature of the characteristics of matter and the way of describing them at higher levels of organization (see Sokołowski 2006b). In this way, for example, the relations between biology and physics should be understood in the context of emergence rather than strict entailment (Sokołowski 2001, p.216). This type of approach to explaining natural phenomena was close to many representatives of the Kraków interdisciplinary tradition, including the Kraków School of Philosophy in Science. It is worth mentioning Fuliński's views here (e.g., Fuliński 1993). Methodological naturalism in Sokołowski's papers is not equivalent to ontological naturalism, which was clearly exposed by Życiński in his works (Życiński 2003).

Sokołowski also had other views in common with Życiński. Like Życiński, he was critical of various manifestations of contemporary ontological reductionism—e.g. in the field of sociobiology, with the representatives of which Życiński strongly argued (Sokołowski 2001, p.219; Życiński 1993, pp.243–268). From the point of view of metaphilosophy, Sokołowski also addressed the problem of "rationality" and today's challenges related to it, and he did

it in a similar way to Życiński. This is especially visible in his criticism of postmodernism. Sokołowski noted that postmodernism—and the phenomena that go hand in hand with it, such as political correctness, which "negates objective truth in the name of higher social reasons" (Sokołowski 2006a, p.379)—constitutes a significant social threat to concept of rationality, which was actually a similar view also for Życiński (Życiński 1994). It can therefore be said that Sokołowski, like Życiński, defended the concept of rationality against two extreme forms—its narrow understanding in positivizing trends (in the second half of the 20th century, the continuators of this line of thinking included, among others, representatives of sociobiology) and the trend that completely rejected it (postmodernism).

4.3 The border problems of science and metaphysics

4.3.1 Mathematical Universe hypothesis

The next issue to be discussed concerns the mathematical nature of the world. Mathematical universe hypothesis was one of the central issues undertaken by various representatives of the Kraków School of Philosophy in Science. This issue was addressed by both philosophers and physicists associated with Heller's milieu (Fuliński et al.). Sokołowski also drew attention to this issue in his papers (see e.g, Sokołowski 1987, 1990, 2011b, 2015a). Many representatives of the Kraków School sympathized with Platonizing trends in the philosophy of mathematics and ontology. Others, such as Fuliński or Sokołowski, had a slightly more nuanced position on this issue. Regardless, Sokolowski believed that the problem itself was important and worth addressing. In turn, his declarations (see above) suggest clear sympathy with Heller's views on the mathematical nature of the world.

In his works, Sokołowski states that the mathematical nature of the world should be considered as the foundation ("initial assumption") of modern natural sciences. Mathematics plays a key role in acquiring knowledge about the world, both at the elementary level and in the case of complex living systems. In one of his paper, Sokołowski answers the question "what does it mean that nature is mathematical?" he replies that "nature as a whole and in each of its parts is subject to the laws of nature, which constitute a mathematical structure, i.e. create a deductive system of mathematical theorems" (Sokołowski 1990). Sokołowski places significant emphasis on the mathematizability of nature (as the possibility of describing the world using mathematical methods) and treats the world as a kind of implementation of a mathematical structure.

Nevertheless, he discusses Platonism and states that in the context of the problem of the existence of mathematics as a "world of ideas" or a "third world" (Popper), it is impossible to ignore important, and at the same time virtually unsolvable, terminological and ontological problems: How do mathematical objects exist? Why do they exist? What is the relationship between the world of mathematics and nature? What is the correspondence between the physical world and the world of mathematics? Sokołowski also draws attention to the issue of "redundancy of mathematics" (mathematics is developing faster than its applications; physics uses only a fragment of existing mathematical knowledge) and, in reference to this, poses the problem of the relationship of mathematics to the world (Sokołowski 2011b, pp.217–220). In his recent works, Sokołowski also draws attention to certain emerging difficulties in connection with the application of the concept of the mathematical nature of the world to higher levels of matter organization (animate beings) (Sokołowski 2015a, p.74).

Sokołowski shows some caution in formulating answers here. Without questioning the idea of the mathematical nature of the world, he points out that we cannot convincingly answer the question of why nature is mathematical. This type of view also appears in Sokołowski's later papers (see: (Sokołowski 2015a, p.65). He believes that the mathematical nature of the universe is an important property of our world, and at the same time—following Einstein—he believed that the natural world is more complex than the possible philosophical answers appearing in the dispute (Sokołowski 1987, pp.190–191; 2015a, p.67). This also leads him to express doubts about Penrose-style Platonism (Sokołowski 2011b, p.218).

On the other hand, Sokołowski is more inclined to adopt the milder position that nature is rational; this thesis is not equal to the stronger thesis that nature is mathematical (Sokołowski 2001, p.215). This approach to the problem allows us to partially overcome the difficulties (which, however, does not eliminate them), and is also in line with the assumptions shared by the Kraków School[7]. It should be noted that this is not a very strong view, therefore it would be interesting to compare Sokołowski's position with, for example, the the idea of the field of rationality from Życiński's approach.

4.3.2 Relationship between science and religion

Another important area of philosophical exploration was the relationship between science and religion. In this area, Sokołowski was mainly interested in methodological aspects. Sokołowski did not deal with controversial issues in detail (such as the "creation-evolution" problem[8]), but rather emphasized the need for a comprehensive modification of religious reflection in the context of developing sciences. He also referred to Christianity, criticizing in particular attempts to renew systemic Christian philosophy, that appeared especially in Catholicism. He also applied his comments on systemic thinking to religion, writing, for example, that "religion does not provide a comprehensive image of the world, but leaves large gaps that can be filled by other conceptual systems, such as empirical sciences or art" (Sokołowski 1989, p.199).

Thus, Sokołowski emphasized the need to develop a new philosophy consistent with Christian doctrine, but it could not be a strictly systemic philosophy like Thomism which for several centuries was considered the most adequate form of Catholic philosophy. "Therefore, there cannot be a 'Christian philosophy' as a doctrine implied by the dogmas of faith, this term can only be used for the systems of Christian philosophers from the past (Augustineism, Thomism). Rather, we should speak of a philosophy consistent with Christianity, and this criterion allows for a number of systems that vary considerably different from each other" (Sokołowski 1989, p.199). Sokołowski was skeptical about attempts to create a new synthesis of science, theology and broadly understood humanistic culture, pointing out that in this respect profound transformations of the current way of thinking about religion would be necessary (Sokołowski 1991, p.265). Despite these remarks, Sokołowski became known as a supporter of the idea of a possible symbiosis of theology and science, suggesting non-contradiction between the credo and scientific knowledge (se e.g., Sokołowski 2001, p.214;

7. However, it is worth saying that in this milieu, rationality was often identified with mathematicality, or there was talk of mathematical-type rationality.

8. It is worth adding that he himself declared his opposition to pseudoscientific concepts based on religious foundation. This is illustrated by an example statement about creationism: "this phenomenon is a typical example of aggressive ignorance drawing its strength from ignorance" (Sokołowski 1991, pp.259–260).

2014, p.180). In science-religion discussions, this places him on the side of accommodationism (a non-confrontational model in McGrath's approach), and the comments on the possible harmonization of science and religion—provided that theology (especially its metaphysical part) undergoes essential transformations—resemble the views of Heller and Życiński (see e.g., Życiński 1990).

In relation to the issue of the relationship between science and theology, Sokołowski eagerly referred to the views of Heller and Życiński, whose ideas he particularly valued (see, Sokołowski 1993; 2014, p.187). Sokołowski believed that natural sciences—especially physics—are able to describe various aspects of the material world, but he remained very cautious about the possibility of formulating a final theory that would express the absolute truth about reality and become to be the end of scientific discovery process (Sokołowski 2011a). Sokołowski showed skepticism when it comes to the possibility of formulating a theory that would coherently describe all physical phenomena at the elementary level. He tried to supplement the scientific image of the world with a religious perspective. He also believed that the institution of the Church faced a serious task related to the need to modify theology in its philosophical layer. Such a change requires taking into account the achievements of modern science. Sokołowski believed that the change in the Church's attitude towards scientific findings cannot be just a top-down reform, but the result of the intellectual development of Catholic society (Sokołowski 1993, p.123).

Both the postulate of opening theology to science and the critical attitude towards systemic philosophy of the Thomistic type—promoted in Polish philosophy especially at the Catholic University of Lublin—placed Sokołowski very close to the views represented by representatives of the Kraków School of Philosophy in Science (see e.g., Trombik 2021, pp.228–229; Trombik and Polak 2022). Sokołowski's views can actually be considered as typical of this philosophical milieu, in which similar ideas were also expressed by philosophizing scientists such as Fuliński (Trombik 2023). His recent comments in the public space indicate that he was discussing the existence of God, arguing on this issue mainly with Jan Woleński (Sokołowski 2024).

5. Summary

The analysis of Sokołowski's articles shows that this physicist has addressed a number of philosophical issues over the course of several dozen years (since the 1980s). The content of these papers proves the author's competences, primarily in the area of philosophy of nature and philosophy of science. Sokołowski discussed key philosophical problems arising in connection with the development of science, especially physics and cosmology.

Sokołowski seems very balanced in his views, avoiding radical philosophical positions. It is clearly visible that—as he declares—he avoids practicing philosophy along the lines of philosophical systems. He takes the findings of modern science as the starting point for his philosophical analyzes and tries to place them in the context of classical problems of philosophy, including metaphysics (e.g. the issue of the mathematical nature of the world). Like Heller or Życinski, Sokołowski recognised the great importance of scientific theories in the process of constructing an image of the world. Therefore, he considered certain philosophical consequences and conducted methodological analyses.

One of the goals of this paper was to place Sokołowski's views against the background of the concept of philosophy in science and the philosophical tradition of the Kraków School of Philosophy in Science. The above reconstruction and analysis of selected philosophical ideas appearing in Sokołowski's works shows that this physicist philosophized in a way consistent with the foundations of philosophy in science (I mean both problematic and methodological convergence, i.e. the application of the theoretical assumptions of Heller's concept). Some of his views (proposals of possible solutions to philosophical problems) also seem to be either similar or at least partially consistent with what was presented in the OBI and the Copernicus Center for Interdisciplinary Studies.

The analyzes conducted indicate that in Sokołowski's case there are similarities with the phenomenon of the Kraków School of Philosophy in Science. These similarities are visible both in the subject matter of the papers (area of expertise), as well as in the way of philosophizing and in the proposals for responding to certain philosophical problems. It is also worth noting here that Sokołowski undertook philosophical analyzes with clear references to the perspective of metaphysics, which is not obvious among modern philosophizing scientists and science popularizers (at least taking into account their declarations). There are many indications that these similarities are not accidental, but are related to the fact that Sokołowski formulated his views in the context of the School's extensive activities. Joint discussions with Heller and other representatives of the School could have had a significant impact on the shape of Sokołowski's philosophy, which fits into the huge interdisciplinary traditions of Krakow.

I think there are reasons to say that Sokołowski philosophized in the same way as representatives of the Kraków School of Philosophy in Science. We can speak here not only of a certain significant similarity, but also of a certain influence of Sokołowski on the milieu of this School. For many years, Sokołowski has been regularly participating in the life of the Kraków interdisciplinary community centered around Heller. Sokołowski has published many papers in books and a periodical closely related to the program of philosophy in science ("Zagadnienia Filozoficzne w Nauce"), published by OBI and currently by the Copernicus Center. Sokołowski is also quoted by Heller and his group of collaborators (e.g. Życiński, Szydłowski, Pabjan, Grygiel, Janusz). All this makes it possible to consider Sokołowski a representative of the Kraków School.

The assessment of Sokołowski's views requires placing them in a historical and philosophical context. Sokołowski's activities—both his publications and his participation in the interdisciplinary milieu through participation in conferences, etc.—were part of the tradition of dialogue between science and philosophy that was conducted in Kraków during the period of political transformation in Poland. Maintaining this dialogue should be assessed positively, especially since Sokołowski's works show real attempts to break down „two cultures" (Snow 1999) and can be considered a local attempt to respond to the growing popularity of the phenomenon of "philosophizing scientists" in the West. It is worth emphasizing here that in a similar period—i.e. from the 1980s—this phenomenon was also gaining importance in Poland, as evidenced by the works of scientists such as Antoni Hoffman, Władysław Kunicki-Goldfinger, Bernard Korzeniewski and others. I think it would be worth examining the formation of this phenomenon in Poland in recent decades, also taking into account Sokołowski's philosophical works. I indicate this as a research perspective worth undertaking to show the importance of Polish interdisciplinary traditions for the native culture. By the

way, it would perhaps show the specificity of Polish philosophy practiced in the context of science.

Sokołowski's various philosophical views also deserve further analyses, comments and polemical discussions. What I mean here is not only a further, more in-depth analysis of these views from the point of view of their historical value, but also about relating them to contemporary problems arising in connection with the development of science and philosophy. Undoubtedly, analyzes of Sokołowski's views in the context of traditional problems of philosophy would also deserve attention. I mean, for example, the issues of reductionism and naturalism (in the ontological version) or the relationship between science and religion. On these issues, Sokołowski represented views that were far from the views expressed in the works of many contemporary philosophizing scientists (especially from the Anglo-American circle, like R. Dawkins, S. Harris etc.), which makes it seem justified to compare and evaluate Sokołowski's views with theirs.

Finally, it seems justified to conduct further research on the phenomenon of the Kraków School of Philosophy in Science. The publications so far constitute a contribution to further analyses, at the same time showing that the research area here is very extensive (taking into account the fact that when I talk about the School, I mean the activities of a very large group of philosophers and scientists from Krakow, whose involvement in the local interdisciplinary milieu includes for almost half a century).

References

Burtyn, Stanisław. 1997. Idea matematyczności przyrody a problem jedności nauk przyrodniczych. *Studia Philosophiae Christianae* 33 (2): 95–101.

Czerniawski, Jan. 2009. *Ruch, przestrzeń, czas. Protofizyczne i metafizyczne aspekty podstaw fizyki relatywistycznej.* Kraków: Wydawnictwo UJ.

Czerniawski, Jan. 2012. Protofizyka a istota teorii względności. *Przegląd Filozoficzny – Nowa Seria* 83 (3): 187–199.

Dąbek, Dariusz. 2016. Nauka i religia w kosmologii Edwarda Artura Milne'a. *Zagadnienia Naukoznawstwa* 52 (2): 275–292.

Davies, Paul. 1996. Zasada antropiczna. *Postępy Fizyki* 37 (3): 213–258.

Filipek, Magdalena. 2008. Elementy absolutne w fizyce w kontekście filozofii Maxa Plancka. *Studia Philosophiae Christianae* 44 (2): 223–237.

Fuliński, Andrzej. 1993. O chaosie i przypadku, a także o determinizmie, redukcjonizmie i innych grzechach fizyków czyli o zmianach w obrazie świata widzianych okiem jednego z nich. *Znak* 465:31–49.

Grygiel, Wojciech. 2010. Teoria superstrun i Lee Smolina kłopoty z fizyką. *Filozofia Nauki* 18 (3): 141–152.

Heller, Michał. 1986. Jak możliwa jest „filozofia w nauce"? *Studia Philosophiae Christianae* 22:7–19.

Heller, Michał. 1990. Nowa fizyka – perspektywy trwającej rewolucji. In *Nauka – religia – dzieje. V seminarium w Castel Gandolfo 8–11 sierpnia 1988,* edited by Jerzy Janik and Piotr Lenartowicz, 66–83. Kraków: Wydział Filozoficzny Towarzystwa Jezusowego.

Heller, Michał. 1992a. *Filozofia nauki: wprowadzenie.* Kraków: Wydawnictwo Naukowe Papieskiej Akademii Teologicznej.

Heller, Michał. 1992b. *Filozofia świata.* Kraków: Znak.

Heller, Michał. 2013. *Filozofia kosmologii.* Kraków: Copernicus Center Press.

Heller, Michał. 2014. *Granice nauki.* Kraków: Copernicus Center Press.

Heller, Michał. 2019. How is philosophy in science possible? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 66:231–249.

Heller, Michał, Zbigniew Liana, Janusz Mączka, Adam Olszewski, and Włodzimierz Skoczny. 1999. Jak filozofuje się w OBI? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 25:20–29.

Hołda, Miłosz. 2014. *Teistyczne podstawy nauki. Epistemologiczne argumenty za istnieniem Boga.* Tarnów: Biblos.

Jacyna-Onyszkiewicz, Zbigniew. 2018. Kosmologia kwantowa a ideologia materialistyczna. In *Wpływ ideologii na naukę i życie społeczne,* edited by M. Rucki, 303–315. Warszawa: Chrześcijańskie Forum Pracowników Nauki.

Janowski, Jarosław. 2016. *Zagadnienie istnienia i natury czasu w wybranych modelach kosmologicznych.* Warszawa: Liberi Libri.

Janusz, Robert. 2017. Stulecie kosmologicznych prac Einsteina i de Sittera. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 63:167–181.

Jodkowski, Kazimierz. 2007. *Spór ewolucjonizmu z kreacjonizmem.* Warszawa: Megas.

Lemańska, Anna. 2020. Mathematicalness or mathematicability of nature? *Studia Philosophiae Christianae* 56:61–80.

Liana, Zbigniew, and Janusz Mączka. 1999. Z kroniki OBI. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 25:133–152.

Pabjan, Tadeusz. 2013. Filozoficzne idee w fizyce i kosmologii Alberta Einsteina. *Filozofia Nauki* 21 (2): 131–143.

Peacocke, Arthur. 1991. *Teologia i nauki przyrodnicze.* Kraków: Znak.

Polak, Paweł. 2011. 19th Century Beginnings of the Kraków Philosophy of Nature. In *Philosophy in Science. Methods and Applications,* edited by B. Brożek, J. Mączka, and W.P. Grygiel, 325–333. Kraków: Copernicus Center Press.

Polak, Paweł. 2018. Tradycja krakowskiej filozofii w nauce: między XIX a XXI wiekiem. In *40 lat filozofii w uczelni papieskiej w Krakowie,* edited by Jarosław Jagiełło, 491–513. Kraków: Wydawnictwo UPJPII.

Polak, Paweł. 2019. Philosophy in science: a name with a long intellectual tradition. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 66:251–270.

Polak, Paweł, and Kamil Trombik. 2022. The Kraków School of Philosophy in Science: profiting from Two Traditions. *Edukacja Filozoficzna* 74:205–229.

Popper, Karl. 1987. Hegel i nowy trybalizm. In *Filozofować w kontekście nauki,* edited by Michał Heller, Alicja Michalik, and Józef Życiński, 19–33. Kraków: Polskie Towarzystwo Teologiczne.

Rodzeń, Jacek. 2005. *Czy sukcesy nauki są cudem? Studium filozoficzno-metodologiczne argumentacji z sukcesu nauki na rzecz realizmu naukowego.* Tarnów: Biblos.

Schrödinger, Erwin. 2017. *Przyroda i Grecy. Nauki przyrodnicze i humanistyczne.* Warszawa: IFiS PAN.

Sierotowicz, Tadeusz. 1997. Realizm w kontekście nauki. *Filozofia Nauki* 17:27–38.

Snow, Charles. 1999. *Dwie kultury.* Warszawa: Prószyński i S-ka.

Sobkowiak, Sebastian. 2019. Relacje między filozofią a współczesną nauką na przykładzie pism Weinera Heisenberga. *Filozofia Chrześcijańska* 16:139–159.

Sokołowski, Leszek. 1978. Czy kosmologia jest nauką empiryczną? *Studia Filozoficzne,* no. 6, 65–71.

Sokołowski, Leszek. 1983. Język i metoda [recenzja książki: j. życiński, Język i metoda]. *Znak* 342-343:1015–1016.

Sokołowski, Leszek. 1984. O Galileuszu, nauce i uprzedzeniach. *Przegląd Powszechny,* 64–75.

Sokołowski, Leszek. 1986. Pluralizm wizji świata. O filozofii nauki Stefana Amsterdamskiego. *Studia Philosophiae Christianae* 22 (2): 197–207.

Sokołowski, Leszek. 1987. Alberta Einsteina filozofia fizyki. In *Filozofować w kontekście nauki,* edited by Michał Heller, Alicja Michalik, and Józef Życiński, 187–201. Kraków: Polskie Towarzystwo Teologiczne.

Sokołowski, Leszek. 1989. Głos w dyskusji po referacie ks. Gadacza. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 11:44–45.

Sokołowski, Leszek. 1990. Nadwyżkowość matematyki. In *Matematyczność przyrody,* edited by Michał Heller, Józef Życiński, and Alicja Michalik, 56–71. Kraków: OBI.

Sokołowski, Leszek. 1991. Posłowie tłumacza. In *Teologia i nauki przyrodnicze,* 257–268. Kraków: Znak.

Sokołowski, Leszek. 1993. Kościół a nauka [recenzja książki: m. Heller, Nowa fizyka i nowa teologia]. *Znak* 456:119–123.

Sokołowski, Leszek. 1994. Wszechświat, jego wymiar i ewolucja. In *Kosmos i filozofia,* edited by Zdzisław Golda and Michał Heller, 29–58. Kraków: OBI.

Sokołowski, Leszek. 1996. W poszukiwaniu teorii ostatecznej. In *Przestrzenie Księdza Cogito. Księdzu Michałowi Hellerowi w sześćdziesiątą rocznicę urodzin,* edited by Stanisław Wszołek, 88–113. Tarnów: Biblos.

Sokołowski, Leszek. 1999. Mała apologia redukcjonizmu. In *Sensy i nonsensy w nauce i filozofii,* edited by Michał Heller, Janusz Mączka, and Jacek Urbaniec, 57–77. Kraków-Tarnów: OBI-Biblos.

Sokołowski, Leszek. 2000. Czas a grawitacja kwantowa. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 27:3–32.

Sokołowski, Leszek. 2001. Współczesne zagrożenia racjonalizmu. *Analecta Cracoviensia* 33:213–232.

Sokołowski, Leszek. 2006a. Alicja w Krainie Czarów, czyli społeczeństwo postindustrialne wobec nauki. In *Wyzwania racjonalności. Księdzu Michałowi Hellerowi współpracownicy i uczniowie,* edited by Stanisław Wszołek and Robert Janusz, 378–402. Kraków: WAM-OBI.

Sokołowski, Leszek. 2006b. Teorie efektywne i emergencja fizycznego obrazu świata. In *Struktura i emergencja,* edited by Michał Heller and Janusz Mączka, 121–139. Kraków-Tarnów: PAU-OBI-Biblos.

Sokołowski, Leszek. 2007a. Człowiek jako twórca teorii fizycznych w wieloświecie. In *Człowiek: twór Wszechświata – twórca nauki,* edited by Michał Heller, Robert Janusz, and Janusz Mączka, 73–86. Kraków-Tarnów: PAU-OBI.

Sokołowski, Leszek. 2007b. Metric gravity theories and cosmology. I. Physical interpretation and viability. *Classical and Quantum Gravity* 24:3391–3411.

Sokołowski, Leszek. 2008a. Czego możemy się nauczyć na przykładzie teorii strun? In *Prawa przyrody,* edited by Michał Heller, Janusz Mączka, Paweł Polak, and Małgorzata Szczerbińska-Polak, 21–41. Kraków-Tarnów: OBI-PAU-UJ-Biblos.

Sokołowski, Leszek. 2008b. Uzasadnianie antropiczne, czyli człowiek we Wszechświecie. In *Prace Komisji Filozofii Nauk PAU. Tom 2,* edited by Jerzy Janik, 87–103. Kraków: PAU.

Sokołowski, Leszek. 2011a. O pewnych podobieństwach filozofii fizyki i religii. In *Czy nauka zastąpi religię?,* edited by Bartosz Brożek and Janusz Mączka, 45–62. Kraków: Copernicus Center Press.

Sokołowski, Leszek. 2011b. Parę uwag o matematyczności przyrody. In *Nauka w filozofii. Oblicza obecności,* edited by Stanisław Burtyn, Małgorzata Czarnocka, Włodzimierz Ługowski, and Anna Michalska, 209–220. Warszawa: IFiS PAN.

Sokołowski, Leszek. 2014. Czy można być chrześcijaninem w wielkim wszechświecie? In *Relacja nauka-wiara. Nowe ujęcie dawnego problemu,* edited by Jacek Golbiak and Monika Hereć, 179–195. Lublin: Wydawnictwo KUL.

Sokołowski, Leszek. 2015a. Co nowego w filozoficznym problemie matematyczności przyrody? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 58:63–88.

Sokołowski, Leszek. 2015b. Granice fizyki w kosmologii. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 59:25–81.

Sokołowski, Leszek. 2015c. Racjonalny chrześcijanin. *Palestra. Pismo Adwokatury Polskiej* 695-696 (11-12): 238–245.

Sokołowski, Leszek. 2017. Kłopoty z jedynością Wszechświata. In *Oblicza filozofii w nauce. Księga pamiątkowa z okazji 80. Urodzin Michała Hellera,* edited by Paweł Polak, Janusz Mączka, and Wojciech Grygiel, 225–253. Kraków: Copernicus Center Press.

Sokołowski, Leszek. 2024. *Polscy uczeni dyskutują o Bogu [głos w dyskusji].* https://wyborcza.pl/magazyn/7,124059,30633489,polscy-uczeni-dyskutuja-o-bogu-jestem-przekonany-jako-fizyk.html. [Online; accessed 2024-02-18]. https://wyborcza.pl/magazyn/7,124059,30633489,polscy-uczeni-dyskutuja-o-bogu-jestem-przekonany-jako-fizyk.html.

Sokołowski, Leszek, and Andrzej Staruszkiewicz. 1987a. Myśl czysta pojmuje rzeczywistość. O filozofii fizyki Alberta Einsteina (I). *Przegląd Powszechny,* no. 2, 176–186.

Sokołowski, Leszek, and Andrzej Staruszkiewicz. 1987b. Myśl czysta pojmuje rzeczywistość. O filozofii fizyki Alberta Einsteina (II). *Przegląd Powszechny,* no. 3, 348–367.

Szydłowski, Marek, and Paweł Tambor. 2010. Prostota modelu kosmologicznego a złożoność wszechświata. *Roczniki Filozoficzne* 58 (2): 153–180.

Szydłowski, Marek, and Paweł Tambor. 2015. Ontologiczne i epistemologiczne aspekty pojęcia „ex nihilo" w modelach kosmologicznych kwantowej kosmogenezy. *Studia Philosophiae Christianae* 51 (1): 141–163.

Szydłowski, Marek, and Paweł Tambor. 2020. Czy fizyczne teorie efektywne są wiarygodną strategią osiągnięcia teorii ostatecznej? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 68:79–116.

Trombik, Kamil. 2019. The origin and development of the Center for Interdisciplinary Studies. A historical outline by 1993. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 66:271–295.

Trombik, Kamil. 2021. *Koncepcje filozofii przyrody w Papieskiej Akademii Teologicznej w Krakowie w latach 1978–1993. Studium historyczno-filozoficzne.* Kraków: Scriptum.

Trombik, Kamil. 2023. Andrzej Fuliński as a representative of the concept of philosophy in science. *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)* 75:233–256.

Trombik, Kamil, and Paweł Polak. 2022. Teologia nauki – propozycja nowego otwarcia teologii na nauki. *Człowiek i Społeczeństwo* 54:49–64.

Turek, Józef. 2005. Filozofia kosmologii – zarys problematyki. *Roczniki Filozoficzne* 53 (2): 269–308.

Wilson, Edward. 2011. *Konsiliencja. Jedność wiedzy.* Poznań: Zysk i S-ka.

Życiński, Józef. 1990. *Trzy kultury.* Poznań: W drodze.

Życiński, Józef. 1993. *Granice racjonalności. Eseje z filozofii nauki.* Warszawa: PWN.

Życiński, Józef. 1994. Postmodernistyczna krytyka racjonalności nauki. *Studia Philosophiae Christianae* 30 (2): 299–312.

Życiński, Józef. 2003. Naturalizm ontologiczny a rola superweniencji w ewolucji biologicznej. *Roczniki Filozoficzne* 51 (3): 7–18.

Życiński, Józef. 2009. *Wszechświat emergentny. Bóg w ewolucji przyrody.* Lublin: Wydawnictwo KUL.

Philosophical Problems in Science (2024), **77**, 71–89
(Zagadnienia Filozoficzne w Nauce)

doi:00000.000

ARTICLE

Orientation in the environment like perceiving affordances? Andrzej Lewicki's account of cognition

Michał Piekarski† and Witold Wachowski*‡

†Cardinal Stefan Wyszyński University in Warsaw
‡Maria Curie-Skłodowska University in Lublin
*Corresponding author. Email: witoldwachowski@gmail.com

Abstract

The purpose of this article is to present Andrzej Lewicki's account of cognition as orientation in the environment, comparing it with James J. Gibson's ecological psychology. To do so, we conduct a comparative analysis of the former's theory of indicators and the latter's theory of affordance. The theoretical frame for our study is cognitive ecology, a research tradition characteristic of various studies of cognition, including contemporary ones. This allows us to show that, despite differences in the backgrounds and methodologies of these researchers, Lewicki can be considered one of the pioneers of contemporary ecological trends in cognitive science, although his influence has not been as widespread as that of Gibson. Our analysis proceeds in several steps. We begin with an overview of the biographies, backgrounds, and interests of the two researchers, as well as a brief introduction to cognitive ecology and related terms. Next, we discuss action/value indicators theory and affordances theory. We then compare Lewicki's and Gibson's approaches in more detail in terms of their use of similar research heuristics. The article ends with conclusions that go beyond historical issues.

Keywords: affordance, cognitive ecology, ecological psychology, indicator of action, indicator of value, methodological individualism, orientation in the environment, tendency of the organism

1. Introduction: two innovative psychologists

The 1950s and 1960s brought about innovative approaches to the issues of cognition, mind, perception, etc. This was when cognitive psychology and cognitive science were created. At that time, two innovative psychological theories were being formed: one in Europe in the Eastern Bloc, authored by Andrzej Lewicki, the other in the United States, authored by James J. Gibson. Their ideas were different in some ways, and at the same time they had much in common, although there was probably no influence between them. Since these researchers conducted their studies at a similar time, they can be considered pioneers of contemporary ecological approaches to research on cognition, although Lewicki's influence

is not as far-reaching as Gibson's. It is worth noting that the former's theory matured some time earlier than the latter's.

Our goal is to discuss Lewicki's cognitive theory in the light of some of the findings of Gibson's ecological psychology. We argue that one can identify similar assumptions and claims in their works. We will demonstrate this by analyzing how they conceptualize the role of the environment in the cognitive activity of the agents living in it. However, since we are primarily concerned here with Lewicki's theory, the framework for our analysis should not be Gibsonian ecological psychology but a more general research tradition that, as we will show, was common to them.

We have focused here on Lewicki's book *The Cognitive Processes and Orientation in the Environment* [orig. *Procesy poznawcze i orientacja w otoczeniu*] (1960), relating it to Gibson's latest book, *The Ecological Approach to Visual Perception* (1979), which includes—as is usually believed—his most mature study of affordance theory. In addition, we refer to supporting literature. Biographical contexts and scientific backgrounds are also important to us, and that is where we start.

A Polish psychologist Andrzej Władysław Tadeusz Lewicki (1910–1972) dealt with experimental psychology, making creative use of research by Ivan P. Pavlov, Jerome S. Bruner, Robert S. Woodworth, and Harold Schlosberg. Lewicki's research is inspired on the one hand by behaviorism and physiologism, and on the other hand by a critical reading of Freudian psychoanalysis and Gestalt psychology of Kurt Koffka and Wolfgang Köhler. He was a founder of the first Polish Department of Clinical Psychology in Poznań. His research covered, among others, the orientation of an agent in the environment. Contrary to Pavlovism that dominated Polish psychology at that time, Lewicki emphasized the active nature of perception and the agent acting in the structured environment, whose influence could not be reduced to a series of stimuli or sensations. Some of his proposals were influenced by Gestalt psychology. He was the author of the original, experimental method of creating artificial concepts. His works include *Mechanism of Distinguishing in the Light of Pavlov's Research* [orig. *Mechanizm odróżniania w świetle nauki Pawłowa*] (1955), *How Educational Difficulties Arise* [orig. *Jak powstają trudności wychowawcze*] (1957), *Outline of Clinical Psychology* [orig. *Psychologia kliniczna w zarysie*] (1968) and the already mentioned *Cognitive Processes and Orientation in the Environment* (1960). The latter work should be considered one of the most important books in Polish post-war psychological literature due to its new, original approach to the cognitive process.

An American psychologist James Jerome Gibson (1904–1979) still influences psychologists, cognitive scientists and more. Over the years, he developed his ecological approach to visual perception, the stages of which are evident in his work, including the books *The Perception of the Visual World* (1950), *The Senses Considered as Perceptual Systems* (1966), and *The Ecological Approach to Visual Perception* (1979). The approach embraces a bold idea of direct perception, unmediated by complex internal data processing using mental representations. Perceptual information, Gibson argues, is present in the environment and is directly perceived, but is independent of our ability to recognize it. The role of the environment has become crucial to his theory of perception.

Gibson's approach has been reconstructed many times, so we will limit its outline here to the necessary minimum, mainly in relation to specific components of Lewicki's approach. However, it is necessary to point out the obvious influence of pragmatism (Edwin B. Holt,

William James), behaviorism (Edward Tolman), phenomenology (Maurice Merleau-Ponty), and Gestalt psychology (Kurt Koffka, Kurt Lewin) on Gibsonian psychology (see Lobo, Heras-Escribano, and Travieso 2018). We will come back to some of these inspirations when presenting the affordance theory. In addition, it is worth pointing to his research during the Second World War, while serving in the Air Force (the research covered, inter alia, the visual identification of aircraft and the impact of training films), his early project of global psychophysics of objects and events, analysis in the field of optics (we will mention later on the so-called ecological optics), as well as the use of research conducted by his wife Eleanor Gibson, a psychologist dealing with perceptual development in children (Gibson 1969; Hochberg 1994; Heft 2001; Lobo, Heras-Escribano, and Travieso 2018).

We will reconstruct Lewicki's idea of cognition as orientation in the environment in more detail (within the thematic scope of the article), due to the shortage of Lewicki's works in English: only one, albeit important, chapter of the book *The Cognitive Processes and Orientation in the Environment* (2016) has been translated.

The observations and theses of Lewicki and Gibson seem to fit very well into the research tradition we call cognitive ecology.[1] In short, cognitive ecology involves the study of cognitive phenomena in their biological, social and cultural contexts, describing the heterogeneous interactions in cognitive ecosystems. In this light, our cultural practices are important components of human cognition. This research tradition developed particularly in the 1990s, but it was present even earlier: pioneers include the biologist Jakob von Uexküll (1921; newer edition, see e.g., 2022) who formulated *Umwelt* ("surround-world") theory at the beginning of the last century, and approaches that preceded the "ecological boom" in cognitive science were, for example, anthropologists Gregory Bateson's concept of ecology of mind (Bateson 1972) and Jean Lave's studies on situated learning (Lave 1988).

The common heuristic in this research tradition is that environmental factors are taken into account not at a later stage of analysis, but at its beginning, which often turns out to be crucial. This heuristic, like others, can be unreliable. However, it can protect against a neurocentric or reductionist approach. Thus, this research tradition moves away from so-called methodological individualism and makes it possible to study cases and types of cognitive processes that are not reducible to an agent's brain activity (see e.g., Robbins and Aydede 2008).

The aforementioned "methodological individualism" in the research on cognition implies, in its most radical form, the approach according to which the study of the individual agent is both necessary and sufficient to know all the important aspects of cognitive processes. Relations and relational properties are ignored here since their nature is not individual (see Heath 2024).

In the following second part of the article, we present and discuss Lewicki's theory of action/value indicators along with his understanding of cognition, as well as Gibson's affordance theory, taking into account their sources and inspirations. The third part provides a relatively detailed analysis of Lewicki's approach in comparison with some of Gibson's findings, using the cognitive ecology framework; we focus here on orientation in the environment in the context of value perception, the comparison of affordances and indicators, and the social and cultural dimension of cognition. In the final, fourth part, we summarize our

1. Partly after Edwin Hutchins, e.g. (2010).

analysis and evaluate Lewicki's approach, not only in terms of view of potential competition with Gibson's, but also of their possible complementarity, which is worth considering for use in today's research.

2. Theory of indicators of action/value and theory of affordance

The understanding of "cognition" that Lewicki presents in *The Cognitive Processes and Orientation in the Environment* (1960) is focused on the agents' orientation in their environment, a kind of cognitive niche in which they try to maintain the balance of the organism and achieve their different goals. In general, Lewicki states that psychology is not so much about a description of mental processes understood in terms of external experience, or their explanation in relation to specific external or internal conditions, as about explaining the behavior of a given organism in relation to these processes (Lewicki 1960, pp.7–18). His starting point is the statement that the current psychology based on the method of introspection, as well as the language used to articulate its results, are unjustified with regard to explaining cognitive processes understood in terms of the mechanisms responsible for behavior. The cause of this, according to Lewicki, is the inadequacy of applying the methods of introspective psychology to behavioral mechanisms that cannot be reduced to mental and consciousness mechanisms. Apart from that, introspection treats psychic processes as non-spatial and therefore does not explain how these may influence the actions of the organism and cannot be applied to animals without committing the error of anthropomorphism.

It should be added that Lewicki appreciated the efforts of Freudian psychoanalysis in moving away from classical psychology towards a new psychological language, but he nevertheless criticized it on the grounds that it was based on introspective methods in its assumptions and was skeptical about explaining the "true nature of psychological phenomena." According to Lewicki (1960, pp.43–44), psychoanalysis—like introspection—is not able to explain how mental processes can influence biological processes carried out by the organism.

Another important point of reference for Lewicki's research was the achievements of the Russian physiologist Ivan Pavlov. On the one hand, Pavlov's research greatly influenced modern behaviorism, especially Watson's, and defined its conceptual resources. On the other, to understand the specificity of Lewicki's work in post-war Poland, especially in the 1940s and 1950s, one must bear in mind that access to the latest psychological literature from Western Europe and the USA was significantly difficult, so that researchers relied mainly on Pavlovism and recent Soviet psychology (cf. Strelau 2010).

Lewicki begins his analyzes with the concept of cognition understood as "nervous reflection" (cf. Adrian 1948; Asratyan and Shingarov 1982). Generally speaking, this term denotes systems of nervous processes reflecting specific external stimuli in such a way that the nervous reflection is broken down into individual, elementary phenomena and elementary activities of the nervous tissue, which remain in specific spatial and temporal relations to one other. Lewicki describes this approach as physiological and contrasts it with a psychological approach, which abstracts from the physiological structure of nervous processes, referring to what has been reflected in the cerebral cortex and to what extent it is consistent with a specific stimulus. He rejected the physiological approach as insufficient for the investigation of animal behavior. Animals mirror the various features of objects as indicators of the objects' values and as indicators of actions to be taken, that is, they "understand" the meaning which these

objects have for them, and it is, according to Lewicki, only this "understanding" that can be named "cognition." The subject of the research is therefore the content of the reflection, not only the governing mechanism. This content, argues Lewicki, cannot be explained with the help of physiological language, but psychological ones that need to be constructed anew (Lewicki 1960, pp.7–18). So cognition not only reflects specific states of the environment, but must also include the "account" of the value of the reflected phenomena and the action that a given organism should perform in a given situation in order to maintain its inner balance (Lewicki 1960, p.186). This means that a living being must in some sense "understand" the environment in which it lives.

According to Lewicki, this "understanding" should be explained. First, due to its ambiguity and fuzziness, he proposes to replace the notion of understanding with the term of orientation in the environment. Semantically, this notion is related to phrases such as "orientation (of rats) in a maze" or "orientation in the woodland of bees returning to the hive." Lewicki states:

> Namely, it must be acknowledged that the basic component of the orientation process is the reflection of the features of an object or—to be more general—the phenomena that are the determinants of the value of the object. I will refer to this component as orientation in value. Orientation in the value of objects found in the environment should be recognized as the principal component of the orientation process because it must happen in each act of adaptation: adaptation to the environment consists in the fact that an animal reacts proportionally to the value a given object has for it, which means that it must somehow perceive this value and somehow orient itself in it. (Lewicki 2016, p.48)

However, it should be remembered, as Lewicki emphasizes, that the mere orientation in the value of a given object is usually not sufficient for an animal to be guided in its behavior, that is, to assimilate a valuable object or protect itself from harm. It is also necessary to act appropriately, aiming at a given value.

As for the notion of value, Lewicki understands it in terms of pragmatism (he was familiar with the philosophy of John Dewey (cf. Lewicki 1960, p.234)), and not in a moral sense. It should be added that the very notion of value Lewicki owes to the work of researchers associated with Pavlov, although he gives up its physiological interpretation. According to Lewicki, what is valuable is what promotes the organism's survival. More precisely, values are related to the properties of objects that make them necessary to maintain the internal balance of an animal. Therefore, the objects that represent a given value for an animal, e.g. are suitable as food, have specific properties, which consequently constitute the indicators of their value. The existence of indicatiors does not mean that the values are mediated. They are given directly to animals because the objects which the animal perceives as valuable are directly given. Lewicki easily distinguished between positive and negative values. If they are positive, they trigger exploratory actions. If they are negative, then the actions are preservative. The latter take place, for example, when an object encountered by an animal poses a threat to it. Preservative actions often appear at the level of defensive reactions that are learned and instinctive. In other cases, however, the actions are driven by appropriate indicators of value.

Orientation in the environment devoid of an action element adapted to a given value, is an artificial phenomenon and is achievable in laboratory conditions in the simplest of situations. Thus, the orientation of an animal—guided by specific tendencies—in the environment consists of both value orientation and orientation in action. These, in turn, are possible thanks to the indicators of value and indicators pointing to action present in the environment. We will discuss the mechanisms of tendencies and indicators in more detail later in this article.

Gibson's ecological psychology, which has made an important contribution to the study of visual perception, focuses on the agent-environment coupling. As William Mace (2005, pp.199–200) emphasizes, Gibsonian idea of including the environment in understanding of mind required significant changes, the importance of which not everyone appreciated, including the reformulation of the concept of the stimulus and the assumed ontology of the environment. It is also worth looking at the entire history of misrepresentations of Gibson—treated as a "distinguished dissident"—in textbook publications, as there has been a constant effort to understand his ideas in the context of the approaches he rejected, especially in the field of psychology (see Costall and Morris 2015).

The author of *Ecological Approach to Visual Perception* (Gibson 1979) shifts the emphasis from the question of what we are equipped with to explore the world to the question of our cognitive relationships with the world and the structure formed on their basis. The founder of ecological psychology challenges the traditional dualism of the subject and the environment, presenting both concepts in relational, not oppositional terms—a strategy visible throughout his main work (1979). Both components—a perceiver and an environment—form a tailored and dynamic structure. Some environmental characteristics are essential for a particular animal—and conversely, the morphology and skills of an animal have developed in relation to the conditions and possibilities of its environment. Each environment is structured in a specific way, including different configurations of substances, surfaces, and a given medium (in the case of human, the gaseous atmosphere). The elements of a structured environment—including objects, events and their arrangements—give structure to the ambient light, which is reflected from them and reaches the perceiver from different directions. This structured ambient light contains some invariants relating to aspects of the environment that are important, significant to animals and directly perceived by them. We are dealing here with information about the possibilities of action, called "affordances" (Gibson 1979, pp.127–143).

This term was coined by Gibson from the verb "to afford." The neologism describes some relational properties of the components of the environment of an agent, or simply the relations between the environment and the agent, which induce her to behave in a certain way, "offering" its usefulness or "warning" against harm: a stone can be used as a hammer, a large stone—to replace a chair, a chair—in self-defense, while fire requires that one should keep away from it. Equally important are affordances between agents, who afford one another not only through behavior, but also social interaction.

The concept of affordance appeared in Gibson's work in 1966 (although reflections anticipating this term can be traced back to his earlier works) and was then developed until *The Ecological Approach to Visual Perception*.

> I mean simply what things furnish, for good or ill. What they afford the observer, after all, depends on their properties. The simplest affordances, as food, for example, or as a predatory enemy, may well be detected without learning by the

> young of some animals, but in general learning is all-important for this kind of perception. The child learns what things are manipulable and how they can be manipulated, what things are hurtful, what things are edible, what things can be put together with other things or put inside other things—and so on without limit. He also learns what objects can be used as the means to obtain a goal, or to make other desirable objects, or to make people do what he wants them to do. (Gibson 1966, p.285)

In his 1979 book, Gibson strongly emphasizes the systemic and relational nature of affordances while opposing their (purely) phenomenal account:

> An important fact about the affordances of the environment is that they are in a sense objective, real, and physical, unlike values and meanings, which are often supposed to be subjective, phenomenal, and mental. [...] It is equally a fact of the environment and a fact of behaviour. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer. (Gibson 1979, p.129)

One should keep in mind how Gibson understands the information, as well as what the ecological optics proposed by him is all about. The information is not transmitted, does not come to the perceiver, but is simply available, actively obtained by them (Gibson 1979, p.307). When it comes to the optics: unlike physical optics, ecological optics is concerned with the available information for perception, so it deals with many-times-reflected light in the medium (illumination) and observation usually made from a moving position. Concepts relevant to ecological optics are variance and invariance as reciprocal to one another, not time and space. Gibson treats the hypothesis of information in ambient light to specify affordances as central in this optics (Gibson 1979, pp.47–64, 307–309).

The concept of the aforementioned invariants plays an important role in Gibsonian ecological optics. They are related, on the one hand, to the motives and needs of observers and, on the other hand, to the substances and surfaces of their environments. Structured stimulation that takes place in such a system contains information about the physical properties of things as well as (presumably, as Gibson writes) about environmental properties. Affordances—neither physical nor phenomenal—are the properties related to the observer, not only to her perspective, but also to the status and role she plays in an ecological niche (Gibson 1979, pp.143; 310–311).

Affordances are characterized by their stability: they do not represent the perceiver's projection on things when something is needed. This is important when trying to make an analogy between his approach and Gestalt psychology (see Koffka, e.g. 1935, Lewin, e.g. 1936; cf. endnote 2). As Gibson (1979, pp.138–139) claims, "the affordance of something does not change as the need of the observer changes." We may or may not perceive or attend to it, according to our needs, but the affordance is invariant and is always there to be perceived. "An affordance is not bestowed upon an object by a need of an observer and his act of perceiving it. The object offers what it does because it is what it is".

On this occasion, it is also worth pointing to the likely influence of some ideas of pragmatism and radical empiricism (see Holt 1914, e.g. Heft 2001; Legg and Hookway 2024) on the concept of affordances ("both physical and psychical").

Gibson's proposal has been referred to for a time as an attack on the poverty of stimulus hypothesis. According to this hypothesis, experience far underdetermines human knowledge: people receive too few stimuli to identify the object of perception; therefore, biological mechanisms are largely responsible for the derived state. This phenomenon was seen as evidence for a universal grammar that enables children to learn a language despite the lack of sufficient information in the statements they hear (see Chomsky 1980). Gibson, however, treated stimulus as rather rich, relational and changeable. He later rejected the classical notion of the stimulus altogether (1979), which was related to his opposition to the assumed poverty of the real world. In line with this assumption, "A very large part of what we experience and believe to belong to the real world is not real. It is purely subjective, a mental projection upon an inherently colourless and meaningless world" (Costall 2012, p.85). The answer to this assumption was to be the term of affordance. According to it, the agent is not reliant on a neutral space including neutral data, but always functions in a structured world, a space of values, i.e. a specific physical, biological and (in the case of humans) cultural system, using guidelines and solutions existing in her environment.

So every affordance is constituted by a special, quite substantial relationship between the agent and her environment. The result is a stable and dynamic system involving the two. It is not surprising, then, that Gibson (1979) borrows the notion of niche from ecologists and suggests that such a niche—in the light of his approach—is a set of affordances or, more precisely, specific affordances for a given animal in its environment. According to Alan Costall, "the concept of affordances marks a fundamental shift in Gibson's 'ecological approach' from a theory of perception towards a more encompassing ecology of agency" (Costall 2012, p.88).

The status and role of the agent plays in an ecological niche are strongly related to the meaning of the objects in the agent's environment. Gibson does not agree with the idea that we must distinguish between variable things as such before we can learn their meaning. The theory of affordances begins with a new understanding of value and meaning. When we are perceiving an affordance, we are not perceiving a value-free physical object with added meaning. We perceive a value-rich ecological object. "Any substance, any surface, any layout has some affordance for benefit or injury to someone. Physics may be value-free, but ecology is not" (Gibson 1979, p.140). The aforementioned influence of pragmatism is also visible in the case of the Gibsonian understanding of value: we mean here, for example, the rejection of the strong dichotomy between fact and value, or the specific reality of values.

According to Gibson, the values of things are perceived immediately and directly. This is possible, due to the fact that the observer perceives the affordances of the object already specified in the stimulus information. Here we are dealing with a break with the theories of the mediation of experience, according to which, Gibson (1979, p.140) writes, "bare sensations had to be clothed with meaning".

At the same time, as Gibson emphasizes, one should not attach great importance to the ontological status of affordances, because what matters is not the way they exist, but the fact of our access to relevant information.

The idea of the perception of the environment as direct perception and, consequently, Gibsonian anti-representationalism, did not convince advocates of representationalism (e.g., Fodor and Pylyshyn 1981), however, it was appreciated not only in embodied cognition or enactivism (Heras-Escribano 2019), but also, perhaps surprisingly, in the latest studies on

affordance-based design, which tended to be based on a more complex, representationalist Donald Norman's approach (see Masoudi et al. 2019).

3. Beyond methodological individualism

We will take a closer look at how Lewicki uses the heuristic of cognitive ecology, and we will highlight what is also characteristic of Gibson. The relationship between agents and their value-rich environments is crucial here. Next, we compare the role of indicators and affordances as closely as possible. We will supplement our considerations with a reflection on the role of socio-cultural factors in the cognitive ecology of both researchers.

3.1 Orientation in the environment and perceiving values

According to Lewicki, orientation in the environment is the attitude an animal takes to specific values present in the environment in response to indicators of value, which are aimed at specific actions, in order to maintain the organism's balance understood as self-regulation, and to achieve its biological and behavioral goals. By self-regulation, Lewicki understands the adaptation of the organism to the environment which means maintaining an internal balance in the conditions the environment offers (Lewicki 1960, p.182). This solution brings Lewicki's approach closer to the approach defended by Michael T. Turvey and colleagues who modified Gibson's approach. Among other things, they show that perception-action cycles, related to the important mutuality claim of Gibsonian psychology, have a direct and deep connection with thermodynamic principles (e.g., Swenson and Turvey 1991).

In the light of Lewicki's approach, the structured environment can present specific values of "reward" and "punishment" to organisms. The objects offered by such an environment are what an animal strives for or avoids. At this point, it is necessary to mention a particular disposition of organisms, which Lewicki calls a "tendency." His study of tendency is an important contribution to the discussion of animal (including human) understanding situated in the environment. The researcher took this term from Pavlov and then modified it. By "tendency," Pavlov understood the steering process regulating the unconditional-reflex mechanism. It was supposed to be responsible for the sensitivity of an animal to specific stimuli and to condition it to specific reactions. The process of shaping the animal was thus understood as the "essential tendency of the organism." These are processes that arise in the central nervous system as a result of an imbalance in the body and guide its actions. Pavlov distinguished between food, sexual, aggressive, research, etc. tendencies and argued that they are characteristic of both animals and humans. This tendency may contribute to preserve individual organisms or the entire species. Lewicki proposed to treat tendencies as a non-epistemic element of cognition, responsible for shaping its positive or negative character. Due to the fact that an animal shows a specific tendency towards the environment, cognition can be oriented towards specific values, or, more precisely, specific valuable objects.

The tendency is understood by Lewicki psychologically and not physiologically, as Pavlov proposed. For Pavlov tendency is not an experience, but a specific kind of nervous processes (Lewicki 1960, p.161). Thus, it concerns a specific aspect of the organism. For Lewicki, the tendency is "a given direction" (in the Latin sense of *tendo* which means stretching the bow by aiming in a given direction, which metaphorically means striving for a given goal) or the attitude of the whole organism to specific stimuli. Strictly speaking, a tendency expresses

a positive or negative attitude of an animal and a human towards particular objects present in the environment.

Lewicki emphasizes that "tendencies" should not be confused with the term "need" used by Kurt Lewin (1936). Lewin understands "needs" as "psychological forces" derived from organic processes but by no means identical to them. In some cases, the tendency understood in this way take a more conscious form and become a desire to react to a specific stimulus in a certain way. According to Lewicki, however, tendency determines a specific state of the organism and is independent of conscious desires or needs (Lewicki 1960, pp.171–172).

"Tendency" is a superior term in relation to such notions in the field of folk psychology as "desire." It is also an integral component of the organism's behavior and as such it is a kind of life activity which aims to search in the constantly changing environment for "essential conditions of existence necessary for the animal" (Lewicki 1960, pp.168–169). This term, according to Lewicki, plays an important role in research on adaptation, because only it allows to explain and understand the direction of behavior "towards" and "from" the object. Hence, a complete explanation of behavior is possible only after describing the tendency that an organism shows in relation to particular objects and properties. This, in turn, is conditioned by specific pre-structuring of objects. In this sense, it must be said that the environment is a specific active-passive pole of cognition. It is active because the animal is oriented towards the environment due to tendencies or specific attitudes; it is passive because the previously structured environment somehow motivates the animal to show a specific tendency.

It should be added that Lewicki also noticed the importance of constant, unchanging relations between agents and their perceived environments, although, of course, he did not use the term "invariants" like Gibson did. Objects appear unchanging regardless of the agent's position and movement. Lewicki emphasizes that bodily movements are closely correlated with visual perception, thanks to which objects are perceived as constant tactile quantities (Lewicki 1960, p.139).

Lewicki points out that objects can represent specific values for an animal, so they can serve as food or shelter. Therefore, these objects are characterized by specific properties which Lewicki describes as indicators of their value. Such indicators may include various properties: chemical (taste, smell), optical (color, shape, size), acoustic (sounds made by the prey that a given species hunts) and so on. This means that animals follow these properties through sensual contact with the environment. Thus, the behavior of an animal is directly related to the guidance of indicators suggesting a positive or negative value of an object (Lewicki, like Gibson, emphasizes the direct perception of values: they are available directly, because agents perceives the indicators—or affordances—of the object already specified in the stimulus information). It can be said that indicators of value structure the environment "for" an animal, making it a suitable ecological niche understood as a valuable environment.

Lewicki considers the behavior of bees as an example of a value indicator. If bees are given different kinds of food (e.g., a bowl of syrup and an empty bowl) marked by appropriate symbols, e.g., a cross and a circle, and the bowl of syrup is marked with the cross symbol, then, if the position of both symbols and the associated contents of the bowl change, it turns out that in such conditions bees are able to learn to use the shape of a cross as an indicator of the value of food, i.e., a bowl of syrup. After a certain period of training, bees begin to visit only the bowl marked with the cross, which means that they can distinguish between

two different indicators, marked with appropriate symbols (Dembowski 1946; as quoted in Lewicki 1960, p.102). This may seem like associative learning, but it is not. Since animals are able to react only to some indicators, or react differently to some indicators than to others, it should be said that they not only "associate" or "receive" these indicators, but also "distinguish" them, i.e., "select" those which are indicators of benefit, "ignore" those that are indifferent to them, or "reject" those that are detrimental to them (however, we do not find this argument to be convincing). Note that Lewicki, like Gibson, does not assume any stimulus–response framework. For this reason, Lewicki links following the indicators with what Gestalt psychology defined as "insight" (*Einsicht*) into a given situation, that is understanding or empathizing (Köhler 1925). Insight understood in this way is the opposite of association. It is directly related to such phenomena as perceiving an object in a new way or combining it with specific actions, but also perceiving an object in a broader, coherent, non-obvious context. However, we will not explore the topic of insight here, because, in our opinion, this element does not consistently affect Lewicki's approach.

Apart from indicators of value, Lewicki distinguishes indicators of action. The latter appear when the indicators of value are not enough to achieve a given goal. For example, in order to get food, monkeys sometimes have to use other objects as tools, which requires taking into account the properties that indicate such an application (Wong 2016). Indicators of action, as Lewicki claims, make it possible to obtain a valuable object or to avoid any harm that may threaten the animal. It should also be emphasized that indicators of action do not have to be separate from indicators of value. It may be that the same properties perform both functions. This is the case, for example, when the smell of food indicates to the dog both the value of food and the direction in which to look for it (Lewicki 1960, p.104).

Therefore, we can conclude that the mechanism of being guided by indicators of action consists in (1) recognizing certain properties of objects that, for example, make them an effective tool in relation to a specific cognitive task or goal (understanding the situation) and (2) selecting an action or a sequence of activities that can be completed using this object in order to solve a task, e.g., get food.

To some extent, indicators "show" the animal how to act; they are a kind of "invitation" to a specific interaction with the environment. Animals are oriented in their ecological niches, which means that they are focused both on expected values and on specific actions that are a way to use these values in a given situation. The environment is therefore a field of action, not a set of static objects bound by constant relations. Erwin Straus (1956), whose work was known to Lewicki, similarly distinguished the environment understood as landscape (animal environment, ecological niche) from geographical space (abstract space analogous to a cartographic map) (see also field theory of Lewin 1951).

The existence of value indicators and performance indicators thus enables effective orientation in the environment. Thanks to the ability of reflecting specific situations or, more precisely, the indicators contained in them, the agent can perform such actions that, in a given situation, allow the problem to be solved. Thus, the entire process of getting to know a given situation comprises reflecting action and value indicators combined with the situation itself. In this approach, the cognitive process is an action directed at the appropriate situation through indicators of values and indicators of action correlated with the attitude and tendency of a given agent. Cognition, therefore, is an active exploration of the environment by an organism directed at specific goals and values related to its environmental situation.

Lewicki's understanding of the relationship between agents' cognitive activity and their value-reach environments shows quite a few similarities to Gibson's approach, including the role of invariants or the distinction between the physical properties of things and the relational properties of the animal's environment. Of course, Lewicki does not construct any equivalent of Gibson's ecological optics. The author of *The Cognitive Processes and Orientation in the Environment* states that there are two main types of orientation processes depending on their relation to the value of the elements present in the environment. One is simpler as it is directly related only to the reflection of the indicators of value, the response to this component being usually innate (for example secretory, motor or combined). The other is more complex because it involves both reflecting indicators of value and reflecting separate indicators of action. In both cases, there is a correlation between the value of the object and the corresponding actions taken by the animal. However, one should bear in mind that, as Lewicki emphasized, even biological values are not the absolute properties of objects, such as, for example, color or smell. Values are relative insofar as the properties of the object are related to those of the organism. They are therefore specific relational properties. Depending on the nature of the organism, one and the same object may present a positive or negative value, or have none and be neutral. For example, meat has value as food to a dog and is indifferent to herbivores; immersion in water for a long time is beneficial for fish, but lethally dangerous for a terrestrial animal, and so on. At the same time, however, the value that a specific object represents for an animal is closely related to the properties of that object that make it either necessary or harmful for the animal.

3.2 Affordances and indicators

We will take a closer look at what seems to be the most important issue that connects both researchers. Is the role of Gibson's affordance notion in some way analogous to Lewicki's notion of indicators? If so, does the latter anticipate the former in some way, or is it simply a more useful account? Could it be that the "affordances" contradicts that of "indicators" or are they complementary?

In this context, it is worth paying attention to what Robert Shaw, Michael T. Turvey and William Mace (1982) have proposed. They introduced the term "effectivity" to refer to the properties of an animal directed to the environment, as opposed to affordances as an environmental property directed to the animal. This provoked a discussion as to whether (and when) it is more beneficial to understand affordances in the sense of Gibson, as referring to both the animal and its environment, or whether the "affordance–effectivity dual" is more useful (see discussion in Dotov, Nie, and Wit 2012; Michaels 2003).

It seems that, in the most essential way, Lewicki's notion of indicators and tendencies resembles the proposal of Shaw and colleagues. It does not break the ecological continuity of the system composed of the agent and the environment, if we consistently treat both indicators and tendencies as significantly related to the agent's ecological niche system.

Does Gibson's understanding of affordances lead to more independence from methodological individualism than the perspective based on tendencies and indicators? It depends on the research task in question. If our goal is not to characterize the individual as cognitively rooted in the environment, but to understand the mechanisms of the cognitive ecosystem in which the individual exists, the notions and definitions proposed by Lewicki may be equally effective, and possibly more understandable.

However, one can consider a possible complementarity of both approaches within one research approach. We will present two such options.

One possibility is that, although the "affordance" itself does not belong to a pair of terms describing the relationship between the agent and her environment, it may be an important element of the conceptual frame that also includes an "indicator of value," an "indicator of action" and a "tendency." At the same time, it would remain a notion that characterizes a property of the agent-environment system, not just the agent's environment. To imagine the application of these terms, let's take as an example possible cognitive studies on team games such as basketball or football (see research on affordances and action selection in sport, presented in Cappuccio 2019). On the one hand, one can study the individual activity of a player: his perception, motor skills, the ability to cooperate in a team. In this case, we could identify the correlations of his tendencies, improved with training and motivation, with the perception of indicators of value and action associated with the current situation on the pitch, i.e. the location and trajectory of the ball movement and the behavior of individual players. On the other hand, the object of the study could be affordances as properties of a distributed socio-cognitive system (and therefore not properties of an environment), provided that appropriate research questions were formulated. It is also worth remembering that affordance need not be a component of a pair. There is not always only one (significant) affordance in a given situation. In laboratory experiments we tend to focus on a single affordance, however, the "single affordance paradigm" is inconsistent with our everyday cognitive experiences, so it should be replaced by the "multiple affordance paradigm" (Wagman, Caputo, and Stoffregen 2016, p.791; Costall 2012).

A second possibility to link the accounts of both researchers is to use the notions of indicators and tendencies to analyze given affordances. This area still lacks widely accepted solutions. We are dealing with very different positions, such as the continuation of Gibson's approach in cognitive psychology (e.g. Chemero 2003), studies on affordances in design (e.g. Norman 1988; Gaver 1991), or attempts to relate this approach to brain research (e.g. Cisek 2007). Perhaps Lewicki's approach could facilitate studies on affordances in the field of the cognitive ecosystem. In the case of the team of players we have mentioned, the notion of affordances would be the superior unit, and the notions of indicators and tendencies would be used for a more efficient analysis of the relationships within this system.

The question of notion clarity is related to the question of the ontological status of what we are trying to describe as affordances or tendencies-indicators and the way they exist. In the case of Gibson, both his attitude to the objective-subjective dichotomy (which the notion of affordance removes) and his rather instrumental attitude to the notion of affordance should be taken into account: the key question is not whether affordances "exist and are real but whether information is available in ambient light for perceiving them" (Gibson 1979, p.140). It is also a question about the ontological status of value (Gibson was looking for an appropriate term to avoid the term "value" as burdened philosophically). In Lewicki's approach, the world is pre-structured in a certain way by the indicators that exist in it. In this sense, indicators of value and action could be treated as ontologically objective, which means that they are independent of the subject in their existence. However, it should be remembered that agents relate to given objects directly through indicators, as long as they show a specific attitude that is biologically and/or culturally constituted. In this light, the indicators are somewhat epistemically subjective. In analogy to the notion of affordance,

any attempt at a strongly subjective account of the indicators (too strong a link between indicators and agents) or their radical objectification (linking them with the environment without accounting for the tendencies of organisms) will be a mistake.

3.3 Social and cultural factors

In the light of cognitive ecology, human social and cultural practices are crucial components of cognitive activity, not just its background. One might therefore ask to what extent society and culture are important for Lewicki, compared to Gibson. Incidentally, it should be noted here that regardless of the individual views of researchers working on the common ground of cognitive ecology, the concept of "culture" is not limited to the human world. The cultural activity and social organization of non-human animals, including insects, has been studied for years (see e.g., Zuk 2011).

Most of Lewicki's work (Lewicki 1960, 2016) included analyzes relating to the cognitive activity of non-human animals. The researcher emphasizes, however, that what significantly distinguishes human from other creatures is the environment in which he lives, also due to the degree of its socio-cultural modification. In the language of ecology: each animal lives in its own ecological niche, so the human niche should also be taken into account. According to Lewicki, it is not possible to consider human in relation to some "abstract" environment common to her or him and, for example, to wild animals. Humans must always be viewed from the perspective of their own environment. Human environment is the social environment constituted by other people and their artifacts. This is an important remark because, in the social environment, human basically deals, on the one hand, only with artificial objects produced by society as such, and on the other, with the requirements of this society, which cannot be reduced to some form of objectiveness. These requirements are based on various types of normative and symbolic products, such as legal or moral codes, formalized rules, but also unwritten rules such as, for example, good manners. The normativity of these requirements is based on the fact that society expects the individual to behave in accordance with these rules, and any behavior that contradicts them will be met with various consequences. Requirements understood in this way constitute natural environmental values for human and acting in accordance with them allows him to strive for the state of constant self-regulation (Lewicki 1960, pp.186–187).

Thus, Lewicki draws attention not only to the biological but also to the social role of the environment as a condition for self-regulation. For example, one of the key social requirements relevant to the self-regulation (internal balance) of an individual is the need for mutual help and cooperation, or at least refraining from harming others. Therefore, it must be stated that "an individual, entering a given society, already finds a system of specific values in it, which he takes over and develops needs directed at these values" (Lewicki 1960, p.189). The process of becoming human (growing up) in this perspective is closely related to socialization. The child, partly imitating its environment, and partly adapting to the prohibitions and orders, gradually creates the appropriate non-biological needs in accordance with the ideals of the environment. Its biological needs are also socialized. For example, the need for food is related to specific dishes (national "cuisine," home dishes and so on), and the way of eating these dishes becomes consistent with the customs of a given social group (for example eating with a knife and fork by Europeans, or eating with chopsticks by the Japanese). In this sense, the process of socialization lasts all human life. This reveals

the active nature of the environment in which human lives. The speech development in children growing up in a given environment plays an important role here; however, issues of linguistic development was not elaborated from the perspective of the theory of indicators (Lewicki 1960, pp.204–205). Accoring to Lewicki, human has a specific attitude, because his or her environment is socially and axiologically structured in a certain way. Deprived of culture, humans develop abnormally and the process of their socialization is disturbed.

Gibson emphasizes the importance of the notion of affordance in explaining social interaction (as we mentioned earlier). He strongly believes in "the power of the notion of affordances in social psychology," which will thereby renew itself by rejecting useless assumptions (Gibson 1979, p.42). He gives an example of how a mailbox works that "affords letter-mailing to a letter-writing human in a community with a postal system" (Gibson 1979, p.130). He also draws attention to the action possibilities afforded by animals to other animals they live with: "as one moves so does the other, the one sequence of action being suited to the other in a kind of behavioral loop. All social interaction is of this sort—sexual, maternal, competitive, cooperative" (Gibson 1979, p.36; see also remarks in Carvalho 2020).

Gibson stressed the continuity between the natural and the cultural. He was opposed to considering culture as something highly specific, as it is only our human creation. Yet even his own statements on the subject may raise some doubts: "There is only one world, however diverse, and all animals live in it, although we human animals have altered it to suit ourselves. [...] We were created by the world we live in" (Gibson 1979, p.130). Why does he believe that the world we have created does not impact us in a feedback loop? How can we be sure that we control it? The fact of the natural origin of its substrates and its fundamental laws does not justify denying that a new quality could arise against which humans will be forced to react anew. Nevertheless, Gibson provides us with interesting analyzes of cultural phenomena such as human displays, and—especially—movie perception (as well as filmmaking). He proves that the mechanisms governing our perception of the real world do not differ significantly from those governing our perception of "moving images" (Gibson 1966, 1979).

Despite some differences between Lewicki and Gibson, it should be appreciated that both of them take into account all possible levels of the human cognitive ecosystem, including social and cultural ones. This is clearly in line with the basic heuristics of cognitive ecology many years before the development of situated cognition approaches.

4. Summary and final remarks

Our goal was to discuss Lewicki's cognitive theory in light of some of Gibson's findings with which it shares certain similarities. The reason for such a comparison was that the work of the Polish psychologist, practically unknown in the world, contains certain valuable assumptions and claims analogous to those for which the author of *The Ecological Approach to Visual Perception* is quite widely known.

The analogies between Lewicki's and Gibson's findings may seem superficial or short-sighted in light of the ecological psychology that underlies the classical notion of affordance. For this reason, we have used a different, more basic and neutral framework, that of cognitive ecology. This perspective has allowed us, we believe, to emphasize the logic of the comparison

made and to connect Lewicki's theory with what seemed and still seems to be alive and useful in research on cognitive agent-environment interaction.

Thus, the point was not to prove that the Polish researcher is the "second" Gibson as there are too many differences between them. However, comparing the starting points of the two researchers we classify here as cognitive ecologists seems quite interesting. Lewicki dealt with experimental clinical psychology. He bases his perspective on contemporary studies on the behaviour and brain. He selectively used Pavlov's findings. Gibson, on the other hand, made creative use of optics, physics, psychophysics, conducted research on visual identification of aircraft and the impact of training films, as well as collaborated with his wife researching perceptual development. He also referred to phenomenology. References to Gestalt psychology, behaviorism and pragmatism seem to be common elements in the training of both researchers. The different starting points of the researchers also account for some "asymmetry" between their theories; there are therefore few analogous components in them—but those that we have found can be considered valuable.

Gibsonian innovative (and, to some, controversial) affordance theory continues to be applied, and on a much larger scale, especially in cognitive science and design field, albeit with some caveats and modifications. On the other hand, the acceptance of ecological psychology as a whole by mainstream psychology has been less successful because its radical perspective has led to interpretive misunderstandings from the beginning. In the case of Lewicki, some of his research and theoretical findings are simply outdated. However, his account of cognition as orientation in the environment and its theoretical elaboration are noteworthy and could perhaps be applied today.

By referring to cognitive ecology, we analyzed how the Polish psychologist departed from methodological individualism in comparison with the findings of the well-known American "distinguished dissident." Let us summarize what we have found from this perspective:

(1) Lewicki treats understanding as orientation in the environment. This idea makes it possible to indicate elements analogous to Gibson's account, including notion of invariants, as well as an emphasis on the role of bodily activity in the perception process. Both researchers draw attention to the agent's environment, which to some extent coincides with the concept of an ecological niche. The environment is a space of possibilities for action and, at the same time, values. Like Gibson, Lewicki believes that an important role in cognition is played by specifically understood values, perceived directly. Agents do not attribute values to the elements of their environment, but live among them, in value-rich niches related to the needs of those agents. We have shown that the optimal theoretical tool in Lewicki's application of the ecological heuristic is the concept of the action/value index, analogous to Gibson's affordance concept.

(2) One can consider which of the approaches is more profitable or, in some sense, "economical." Gibson's approach appears to be less complex. When we speak of affordance, we refer to the agent and the environment simultaneously. Such an approach to affordances, however, is counterintuitive and leads to distortions and misunderstandings. As for Lewicki's approach, it is more semantically and epistemically transparent. The notion of indicators and its associated tendencies, which form the heart of the notion of environmental orientation, are intuitively less objectionable than Gibson's. Lewicki's approach does not include a developed position in the dispute about mental representations, but a new meaning of the term

"understanding." Lewicki's ecological "philosophy of perception" is not as comprehensive or as elaborately developed as Gibson's.

(3) As we have suggested, comparing the two approaches need not be viewed in terms of competition. On the one hand, it shows how such positions can be elaborated within the tradition of cognitive ecology (including an attempt at a naturalistic approach to values) starting from points as different as brain research, findings in optics and analysis of animal activity in the environment. On the other hand, it is possible to imagine that the notions of affordance and the tendency–indicator could be used complementarily in order to solve different research tasks within one ecological approach, or as compiled into one coherent account, where the notion of affordances is the superior unit, and the notions of indicators and tendencies are used for a more detailed analysis of the relations in a cognitive ecosystem.

(4) Approaches such as the theory of indicators of action/value or the affordance theory show that treating environmental factors as crucial in studies on cognition appeared long before embodied and situated cognition in cognitive science. The case of the author of *The Cognitive Processes and Orientation in the Environment* shows that Gibson was not alone at the time. Although there are few similarities, the presentation of selected findings of Lewicki in the context of Gibson's work seems to be an interesting point of reference both for historical research and for attempts to further use Lewicki's theory or to refine the concept of affordance in some research tasks.

One can imagine that prospective readers might be discouraged by the still weak presence of Lewicki's work in English and its partial obsolescence. However, we have no doubt that both the history of research on cognitive agent-environment interaction and the repertoire of theoretical tools to be used in this field have been enriched.

References

Adrian, Edgar Douglas. 1948. *O fizycznym podłożu wrażeń zmysłowych [About physical basis of sensual impresions].* Translated by Aniela Szwejcerowa-Gruszczyńska and Jerzy Konorski. Biblioteka Wiedzy Współczesnej 1. Warszawa: Spółdzielnia Wydawnicza "Książka".

Asratyan, E. A., and G. Kh. Shingarov. 1982. Lenin's theory of reflection and Pavlov's teaching on higher nervous activity. *Neuroscience and Behavioral Physiology* 12 (4): 357–363. https://doi.org/10.1007/BF01183098.

Bateson, G. 1972. *Steps to an Ecology of Mind.* Chicago: University of Chicago Press. https://www.rauterberg.employee.id.tue.nl/lecturenotes/DDM110%20CAS/Bateson-1972%20Steps%20to%20an%20ecology%20of%20mind.pdf accessed January 5, 2025.

Cappuccio, Massimiliano. 2019. *Handbook of Embodied Cognition and Sport Psychology.* Cambridge, MA: The MIT Press.

Carvalho, Eros Moreira de. 2020. Social Affordance. In *Encyclopedia of Animal Cognition and Behavior,* edited by Jennifer Vonk and Todd Shackelford, 1–4. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47829-6_1870-1.

Chemero, Anthony. 2003. An Outline of a Theory of Affordances. *Ecological Psychology* 15 (2): 181–195. https://doi.org/10.1207/S15326969ECO1502_5.

Chomsky, N. 1980. On Cognitive Structures and their Development: A reply to Piaget. In *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky,* edited by Massimo Piattelli-Palmarini, 35–54. Cambridge, MA: Harvard University Press. https://archive.org/details/languagelearning0000unse_x3z5/ accessed January 4, 2025.

Cisek, Paul. 2007. Cortical mechanisms of action selection: the affordance competition hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362 (1485): 1585–1599. https://doi.org/10.1098/rstb.2007.2054.

Costall, Alan. 2012. Canonical affordances in context. *Avant* 3 (2/2012): 85–93. http://avant.edu.pl/wp-content/uploads/AC-Canonical-affordances-in-context.pdf accessed January 4, 2025.

Costall, Alan, and Paul Morris. 2015. The "textbook Gibson": The assimilation of dissidence. *History of Psychology* 18 (1): 1–14. https://doi.org/10.1037/a0038398.

Dembowski, Jan. 1946. *Psychologia zwierząt [Animal Psychology].* Warszawa: Czytelnik.

Dotov, Dobromir G, Lin Nie, and Matthieu M de Wit. 2012. Understanding affordances: history and contemporary development of Gibson's central concept. *AVANT,* no. 2, 282–295. https://avant.edu.pl/wp-content/uploads/DDLNMW-Understanding-affordances.pdf.

Fodor, J.A., and Z.W. Pylyshyn. 1981. How direct is visual perception?: Some reflections on Gibson's "ecological approach". *Cognition* 9 (2): 139–196. https://doi.org/10.1016/0010-0277(81)90009-3.

Gaver, William W. 1991. Technology affordances. In *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91,* 79–84. New Orleans, Louisiana, United States: ACM Press. https://doi.org/10.1145/108844.108856.

Gibson, Eleanor Jack. 1969. *Principles of Perceptual Learning and Development.* Century psychology series. New York: Appleton-Century-Crofts.

Gibson, James J. 1950. *The Perception of the Visual World.* Cambridge, MA: Riverside Press.

Gibson, James J. 1966. *The Senses Considered as Perceptual Systems.* Boston, MA: Houghton Mifflin.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception.* Boston, MA: Houghton, Mifflin / Company.

Heath, Joseph. 2024. Methodological Individualism. In *The Stanford Encyclopedia of Philosophy,* Summer 2024, edited by Edward N. Zalta and Uri Nodelman. Stanford, CA: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2024/entries/methodological-individualism/.

Heft, Harry. 2001. *Ecological Psychology in Context: James Gibson, Roger Barker, and the Legacy of William James's Radical Empiricism.* New York: Psychology Press. https://doi.org/10.4324/9781410600479.

Heras-Escribano, Manuel. 2019. *The Philosophy of Affordances.* London: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-98830-6.

Hochberg, Julian. 1994. James Jerome Gibson 1904–1979. In *Biographical Memoirs: Volume 63,* 150–171. Washington: National Academy Press. https://doi.org/10.17226/4560.

Holt, Edwin B. (Edwin Bissell). 1914. *The Concept of Consciousness.* London: George Allen & Company. http://archive.org/details/conceptofconscio00holt accessed January 4, 2025.

Hutchins, Edwin. 2010. Cognitive Ecology. *Topics in Cognitive Science* 2 (4): 705–715. https://doi.org/10.1111/j.1756-8765.2010.01089.x.

Köhler, Wolfgang. 1925. *The Mentality of Apes.* Translated by Ella Winter. London: Kegan Paul & Co. http://archive.org/details/in.ernet.dli.2015.187610 accessed January 4, 2025.

Lave, Jean. 1988. *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511609268.

Legg, Catherine, and Christopher Hookway. 2024. Pragmatism. In *The Stanford Encyclopedia of Philosophy,* Winter 2024, edited by Edward N. Zalta and Uri Nodelman. Stanford, CA: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2024/entries/pragmatism/.

Lewicki, Andrzej. 1960. *Procesy poznawcze i orientacja w otoczeniu [The Cognitive Processes and Orientation in the Environment].* 1st ed. Warszawa: Państwowe Wydawnictwo Naukowe.

Lewicki, Andrzej. 2016. Cognition as Orientation in the Environment. Translated by Magdalena Kopczyńska. *AVANT. The Journal of the Philosophical-Interdisciplinary Vanguard* VII (3): 46–67. https://doi.org/10.26913/70302016.0109.0004.

Lewin, Kurt. 1936. *Principles of Topological Psychology.* Translated by Fritz Heider and Grace M. Heider. New York: McGraw-Hill. https://doi.org/10.1037/10019-000.

Lewin, Kurt. 1951. *Field Theory in Social Science.* New York: Harper & Brothers. http://archive.org/details/fieldtheoryinsoc0000kurt accessed January 4, 2025.

Lobo, Lorena, Manuel Heras-Escribano, and David Travieso. 2018. The History and Philosophy of Ecological Psychology. *Frontiers in Psychology* 9:2228. https://doi.org/10.3389/fpsyg.2018.02228.

Mace, William M. 2005. James J. Gibson's Ecological Approach: Perceiving What Exists. *Ethics & the Environment* 10 (2): 195–216. https://doi.org/10.1353/een.2005.0021.

Masoudi, Nafiseh, Georges M. Fadel, Christopher C. Pagano, and Maria Vittoria Elena. 2019. A Review of Affordances and Affordance-Based Design to Address Usability. *Proceedings of the Design Society: International Conference on Engineering Design* 1 (1): 1353–1362. https://doi.org/10.1017/dsi.2019.141.

Michaels, Claire F. 2003. Affordances: Four Points of Debate. *Ecological Psychology* 15 (2): 135–148. https://doi.org/10.1207/S15326969ECO1502_3.

Norman, Donald A. 1988. *The Psychology of Everyday Things.* New York: Basic Books.

Robbins, Philip, and Murat Aydede, eds. 2008. *The Cambridge Handbook of Situated Cognition.* 1st ed. Cambridge University Press. https://doi.org/10.1017/CBO9780511816826.

Shaw, Robert, M. T. Turvey, and William Mace. 1982. Ecological Psychology: The Consequence of a Commitment to Realism. In *Cognition and the Symbolic Processes,* edited by Walter B. Weimer and David S. Palermo, 159–226. Mahwah: Lawrence Erlbaum Associates.

Straus, Erwin. 1956. *Vom Sinn der Sinne. Ein Beitrag zur Grundlegung der Psychologie [From the sense of the senses. Contribution to the foundation of psychology].* Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-66976-7.

Strelau, Jan. 2010. Panorama psychologii w Polsce po II wojnie światowej, ze szczególnym akcentem na pierwsze dekady okresu powojennego [The panorama of psychology in Poland after the Second World War: emphasis on the first decade of the post-war period]. *Czasopismo Psychologiczne* 16 (1): 7–19. http://www.czasopismo psychologiczne.pl/files/articles/2010-16-panorama-psychologii-w-polsce-po-ii-wojnie-wiatowej_-ze-szczeglnym-akcentem-na-pierwsze-dekady-okresu-powojennego.pdf accessed January 4, 2025.

Swenson, R., and M.T. Turvey. 1991. Thermodynamic Reasons for Perception–Action Cycles. *Ecological Psychology* 3 (4): 317–348. https://doi.org/10.1207/s15326969eco0304_2.

Uexküll, Jakob von. 1921. *Umwelt und Innenwelt der Tiere.* Berlin, Heidelberg: Springer.

Uexküll, Jakob von. 2022. *Umwelt und Innenwelt der Tiere.* London: Legare Street Press.

Wagman, Jeffrey B., Sarah E. Caputo, and Thomas A. Stoffregen. 2016. Hierarchical nesting of affordances in a tool use task. *Journal of Experimental Psychology: Human Perception and Performance* 42 (10): 1627–1642. https://doi.org/10.1037/xhp0000251.

Wong, Kate. 2016. Whose Tools Are These? *Scientific American* 316 (1): 10–12. https://doi.org/10.1038/scientificamerican0117-10.

Zuk, M. 2011. *Sex on Six Legs: Lessons on Life, Love, and Language from the Insect World.* New York: Houghton Mifflin Harcourt.