OPEN ACCESS

COGNITIO
Revista de Filosofia
Centro de Estudos de Pragmatismo
São Paulo, v. 25, n. 1, p. 1-15, jan.-dez. 2024
e-ISSN: 2316-5278

https://doi.org/10.23925/2316-5278.2021v22i1:e68263

Automating discovery: what can we learn from the study of abductive reasoning?

Automatizando a descoberta: o que podemos aprender com o estudo do raciocínio abdutivo?

Mariana Vitti Rodrigues[*]
mvittirodrigues@gmail.com

Maria Eunice Quilici Gonzalez[**]
eunice.gonzalez@unesp.br

Abstract: The aim of this paper is to investigate the extent to which abductive inference can be automated. In order to do so, we present the Peircean account of abduction, according to which abduction is the process of generating and selecting an explanatory hypothesis that guides future inquiry (CP 5.171; 1903). Then, we introduce the contemporary concept of abduction characterized as Inference to the Best Explanation (IBE), which aim is to select a hypothesis, among a set of available hypotheses, considering their explanatory potential in terms of likelihood and loveliness (Lipton, 2004). Subsequently, we discuss IBE in relation to Bayesianism, according to which rational agents update their degrees of beliefs in a proposition based on new evidence and explanatory considerations. To illustrate our analysis, we present the software called AI-Descartes, an open-source AI system that combines logical reasoning with symbolic regression, aiming to derive scientific discovery from axiomatic knowledge and experimental data (Cornelio et al., 2023). Finally, we provide considerations about the relevance of studying abduction in the context of Artificial Intelligence.

Keywords: Abduction. AI-Descartes. Automation. Inference to the best explanation.

Recebido em: 09/09/2024.
Aprovado em: 30/09/2024.
Publicado em: 05/12/2024.

Resumo: *O objetivo deste artigo é investigar em que medida a inferência abdutiva pode ser automatizada. Para isso, apresentamos a noção peirceana de abdução, segundo a qual a abdução é o processo de geração e seleção de hipóteses explicativas que orientam a investigação científica (CP 5.171; 1903). Em seguida, apresentamos o conceito contemporâneo de abdução, caracterizado como Inferência à Melhor Explicação (IME), cujo objetivo é selecionar uma hipótese, entre um conjunto de hipóteses disponíveis, considerando seu potencial explicativo em termos de probabilidade e uberdade (Lipton, 2004). Subsequentemente, discutimos IME em relação ao Bayesianismo, segundo o qual agentes racionais atualizam seus graus de crença em uma proposição com base em novas evidências e considerações explicativas. Para ilustrar nossa análise, apresentamos o software denominado AI-Descartes, um sistema de Inteligência Artificial de código aberto que combina raciocínio lógico com regressão simbólica, projetado para derivar descobertas científicas a partir de conhecimento axiomático e dados experimentais (Cornelio et al., 2023). Por fim, apresentamos considerações sobre a relevância do estudo da abdução no contexto da Inteligência Artificial.*

Palavras-chave: *Abdução. AI-Descartes. Automação. Inferência à melhor explicação.*

[*] Universidade Estadual Paulista "Júlio de Mesquita Filho".
[**] Universidade Estadual Paulista "Júlio de Mesquita Filho".

1 Introduction

Can abductive inferences be automated? The objective of this paper is to investigate the concept of abductive inference in the context of the growing automation of scientific discovery. In the history of Artificial Intelligence, attempts to develop algorithmic systems that promote scientific discovery have always received special attention, from DENDRAL in the 60s

(Lindsey, 1993) to AI-Descartes nowadays (Cornelio, 2023). Emphasis will be given to hypotheses on the role of *explanatory inference* in scientific discovery and the extent to which it might be automated. We offer a minimalistic characterization of discovery as a *process involving new findings in relation to previous background knowledge that have the potential to increase human understanding*. Examples of discovery could be the detection of patterns or outliers in massive datasets, or the result of an experiment that confirms a hypothesis promoting understanding in a given research environment. There are cases, however, in which to find a new pattern or correlation seems to create a state of surprise and doubt, instead of directly increasing understanding. This paper focuses on this sort of discovery or, as we are proposing here, discovery through abductive reasoning that involves surprise.

The concept of abduction was introduced by Charles S. Peirce as a form of reasoning that starts with the perception of something to be explained, and ends, provisionally, with the adoption of an explanatory hypothesis. In the modern sense of the word, abduction is characterized as Inference to the Best Explanation (IBE), where the aim is to select a hypothesis based on explanatory considerations (Harman, 1965; Lipton, 2004). The idea of IBE is also discussed in relation to Bayesianism, according to which rational agents update their degrees of beliefs in a proposition, based on new evidence (Bird, 2017; Niiniluoto, 2022; Feldbacher-Escamilla; Gebharter, 2019; Douven, 2022). However difficult it is to explain what exactly abduction is, attempts to automate abduction in the contemporary developments of Artificial Intelligence, and scientific discovery, signal the relevance of investigating the possible perspectives and potential challenges we might face ahead.

To illustrate our analysis, we present the software called AI-Descartes, an open-source AI system that combines logical reasoning with symbolic regression, aiming to derive scientific discovery from axiomatic knowledge and experimental data (Cornelio et al., 2023). Finally, we provide considerations about the relevance of studying abduction in the context of Artificial Intelligence.

2 Peircean notion of abductive reasoning

Charles S. Peirce (1839-1914) investigates the concept of abduction as the core concept of his pragmatism. Different from deduction and induction, the American philosopher, in his later writings, advocates that abduction is the only form of inference to introduce new ideas: "Abduction is the process of forming an explanatory hypothesis. It is the only logical operation which introduces any new idea; for induction does nothing but determine a value, and deduction merely evolves the necessary consequences of a pure hypothesis" (CP 5.171; 1903).

Peirce characterizes inference as "a belief [that] is generated from other beliefs" (W 3:60). Inspired by Bain (1872), he understands belief as something "that upon which a [hu]man is prepared to act" (CP 5.12), thus "belief does not make us act at once, but puts us into such a condition that we shall behave in some certain way, when the occasion arises" (CP 5.373). Inferences, thus characterized, can be subject (or not) to self-control depending on the degree of autonomy agents might have. According to Peirce, abduction, an originative form of reasoning, is considered a form of inference grounded on rational instinct understood as "spontaneous conjectures of creative reason" (Santaella, 2005, p. 189).

The well-known syllogistic form of Peircean abduction is described as follows:

> The surprising fact, C, is observed;
> But if A were true, C would be a matter of course,
> Hence, there is reason to suspect that A is true.
> (CP 5.189)

From the above citation, one would wonder how a machine could possibly be surprised by observing C.

Peirce continues: "Thus, A cannot be abductively inferred, or if you prefer the expression, cannot be abductively conjectured until its entire content is already present in the premise, 'If A were true, C would be a matter of course'" (CP 5.189). Philosophers have challenged this notion of abduction by pointing out that Peirce, in this formal structure, commits the fallacy of affirming the consequent. The abduced hypothesis, in the syllogism, is already present in the minor premise. As an answer to this objection, Peirce scholars have argued, for example, that the syllogistic form of abduction does not necessarily imply a linear temporal commitment. As Anderson (1986, p.157) emphasizes "on Peirce's view it is possible for the hypothesis and its abductive application to occur together. Therefore, abductions may be insightful and originative and still have a logical form". Even if we agree that the abduced hypothesis is already present in the second premise, this syllogistic formulation does not make explicit the inferential steps involved in the generation of explanatory hypotheses.

By the end of his life, Peirce ([1913] 1998, EP2: 463-474) appeals to the human instinct and our ability to guess right, to explain the generation of new hypotheses, going as far as to state that reason is nothing but one of our inborn instincts. Peirce's argument is that humankind did not have enough time to proceed blindly or through trial-and-error, to consolidate science as we know it:

> Think of what trillions of trillions of hypotheses might be made of which one only is true; and yet after two or three or at the very most a dozen guesses, the physicist hits pretty nearly on the correct hypothesis. By chance [s]he would not have been likely to do so in the whole time that has elapsed since the earth was solidified. (CP 5.172).

Peirce also inquires that, if we can see other animals acting and thinking by instinct, why should it be denied to ourselves? In his paper from 1913, entitled "An essay toward improving our reasoning in security and in uberty", the author states that reason is nothing but a part of our instinctive abilities, which he calls reasoning-power or ratiocination:

> *Reasoning-power,* or *Ratiocination*, called by some *Dianoetic Reason*, is the power of drawing inferences that tend toward the truth, when their premises or the virtual assertions from which they set out are true. I regard this power as the principal of human intellectual instincts; and in this statement I select the appellation "instinct" in order to profess my belief that the reasoning-power is related to human nature very much as the wonderful instincts of ants, wasps, etc., are related to their several natures. (Peirce, 1913, p. 464).

In the same paper, Peirce emphasizes that abduction can be a valid argument even if it is a weak (i.e., not deductively truth-preserving) form of argument, because its validity does not depend on its strength. Peirce states that the weakness of abduction guarantees its uberty (or fruitfulness), allowing it to be the reasoning that introduces new ideas to the process of scientific investigation. Ibri explains:

> The ascription of an instinctive aptitude for guessing truths, - apparently an exotic line of argument, - is nothing more than referring to an evolutionary consequence of a kind of attunement of the human mind with nature that enables man, amid an infinity of possible conjectures, to select a given few, among which one proves to be true. (Ibri, 2006, p. 96).

The process of generating new hypotheses, according to this interpretation, is possible by means of organism-nature co-evolution. As part of our instinctive nature, we are able to correctly conjecture hypotheses aimed at explaining unknown phenomena. The appeal to our instinctual abilities to guess right, however, might prevent scholars from further investigating abductive reasoning, through Peircean lenses, as it is hard to combine inference and instinct (some exceptions are Anderson, 1986; Campos,

2009; Hintikka, 1998; Paavola, 2012; Minnameier, 2017; and Bellucci, 2018). Santaella (2005, p. 184) emphasizes that Peirce, by combining instinct and inference, had "[…] the revolutionary and controversial idea of a type of reasoning which is at the same time logical and instinctive, as if there were a logical form for instinct". Thus, by advocating that abduction is a mode of inference, even if it "depends upon altogether different principles" (CP 6.525), i.e. instead of being truth-preserving, is hypothesis-generative, Peirce allows an instinctive logical approach to the process of discovery.

Although we agree that guesses are part of the generation of explanatory hypotheses, from this approach it could be difficult to describe in detail the inferential steps that culminate in those guesses. The question that remains is of the type 'where did the hypothesis **A** come from?'. This kind of imaginary creative constructive generation in the mind of an expert, which proposes an explanatory hypothesis **A**, is the core of abduction. Thus, we may challenge abduction as setting-up syllogism that goes, pseudo-deductively, from premises to conclusion. If the logic of discovery should mirror cognition in the inquiring minds of scientists, the classical inferential mechanisms seem not to be enough. The mind's cognitive use of suggestions, metaphors, intuitions, weird ideas, etc. needs to be further considered.

More recent approaches to the Peircean concept of abduction focus on the role of diagrammatic reasoning in the generation of plausible explanatory hypotheses. The practice of creating imaginative scenarios through diagrammatic experimentation allows the reasoner to anticipate plausible answers to questions in a recursive process of inquiring and guessing (cf. Paavola, 2011; Pietarinen; Bellucci, 2016; Bellucci; Pietarinen, 2020). A diagram can be understood as representing the relationship between the parts of its object (*NEM* IV: 353, 893; *NEM* IV: 275-276, ca. 1895), as the map represents the relationship of the parts of the territory. As a representation, a diagram does not impose restrictions over the flow of imagination, allowing the reasoner the possibility to create, observe, experiment, and manipulate the representation of the object to imagine different scenarios by means of explicating - make it explicit, unfolding, discovering - relations that were formerly implicit (Stjernfelt, 2007, p. 91). Referring to diagrams, Peirce explains that "[…] a very extraordinary feature of Diagrams is that they *show* […] that a consequence does follow, and more marvelous yet, that it *would* follow under all varieties of circumstances accompanying the premises" (NEM IV 317-318, 1909). Pietarinen and Bellucci (2016, p. 474, authors' highlight) emphasize that "[t]he icon-imagination, and the iconic-imaginative moment in reasoning depend on the possibility of *directing* the construction of a perceptual experience".

If we agree that abductive inference is a deliberate, self-controlled and self-corrected process of generation and adoption of explanatory hypotheses (EP2:188, 1903), one can understand that abductive inference occurs when the reasoner can exert some criticism over perception in entertaining with diagrams. So, in abduction, the reasoner is able to adopt an explanatory hypothesis given in the experimentation of imaginary scenarios via diagrammatic manipulation. Thus, by exerting self-control over thought-processes,[1] the reasoner might choose reasonable hypotheses that, if true, would explain the anomalous fact. Hence, the role of abductive reasoning is not to generate an explanatory hypothesis out of nowhere, but to discover a good explanation by diagrammatic reasoning.

According to the above summary of Peirce's abductive reasoning, our answer to our initial question - can abductive inference be automated? - is that abduction thus understood cannot be automated. In general terms, automation can be characterized as the ability to perform tasks based on rules or laws without the continuous supervision of a controlling center, whose objective is the execution of processes

1 According to Peirce, "there are [...] modes of self-control which seem quite instinctive. Next, there is a kind of self-control which results from training. Next, a [hu]man can be his own training-master and thus control his self-control. When this point is reached much or all the training may be conducted in imagination. When a man trains himself, thus controlling control, he must have some moral rule in view, however special and irrational it may be. But next he may undertake to improve this rule; that is, to exercise a control over his control of control. To do this he must have in view something higher than an irrational rule. He must have *some sort of moral principle*. This, in turn, may be controlled by reference to an esthetic ideal of what is fine. There are certainly more grades than I have enumerated. Perhaps their number is indefinite. The brutes are certainly capable of more than one grade of control; but it seems to me that our superiority to them is more due to our greater number of grades of self-control than it is to our versatility" (CP 5.533, our highlights).

that combine programmed commands and feedback control (Groover, 2020). We advocate that the degree of control, control over the ability to control, and self-control over reasoning (cf. CP 5.533), differentiate an automatic disposition to act, from an autonomous form of conducting inference.

Peirce, in his text 'Logical Machines' argues that reasoning machines have inherently two inabilities: "Every reasoning machine […] is destitute of all originality, of all initiative. It cannot find its own problems; it cannot feed itself. It cannot direct itself between different possible procedures." And continues "[...] the capacity of a machine has absolute limitations; it has been contrived to do certain thing, and it cannot do nothing else" (W:6, p. 70, Logical Machines). Although nowadays there are some efforts in the direction of envisioning machines that could have the impetus of originality (Veale et al., 2019), we agree with Peirce that there is no such machine which can be the initiator of its own creativity. Most important, we believe that the Artificial Intelligence community should rethink the willingness to build such a creative machine. Stjernfelt, commenting upon the advances of Large Language Models through Peirce lenses, emphasizes that "Peirce the pragmatist, of course, would highlight the computer´s inability to act upon itself and its environment" (Stjernfelt, 2024, p. 108). In this sense, the possibility to exert self-control, self-correction and self-criticism over reasoning becomes a key criterion that distinguishes, so far, humans from machine thinking.

Abductive reasoning is a cognitive process that requires not-yet formalizable processes such as the feeling of surprise, some deliberate control over action, and the recognition of something as requiring explanation, i.e., as worthy of further inquiry (Peirce to Welby, July 16, 1905, RL 463 *apud* Bellucci 2018, p. 6). It can also be experienced in the form of insight, as an *aha* experience, which requires emotional states typical of living organisms. The instinctual ability to guess correctly also makes it difficult to automate abduction: how could we ascribe instinctual abilities to machines as a form of reasoning-power or reasonableness? Although machines could potentially generate trillions of trillions of explanatory hypotheses, they do not understand why the potential hypotheses could explain the fact in need of an explanation; they would also not recognize something that has to be explained as an embodied and embedded agent. It requires criteria of relevance that are highly context dependent. The adoption of an explanatory hypothesis requires the recognition of a problem to start with. Furthermore, the context-dependency of abductive reasoning makes it difficult to be automated, as it would require establishment of well-structured knowledge domains coupled with generative algorithms and well-defined criteria for hypothesis selection.

In the next section, we explore the concept of abduction as inference to the best explanation, aiming to answer the extent to which abduction can be automated.

3 Abduction as inference to the best explanation

In contemporary philosophy of science, the concept of abduction acquires a new scholarly focus by being primarily characterized as Inference to the Best Explanation (IBE). Generally speaking, IBE aims at the attribution of truth to a given hypothesis based on explanatory considerations (Harman, 1965; Josephson; Josephson, 1994; Lipton, 2004). According to Harman (1965), in IBE "[…] one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there would be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making an inference" (Harman, 1965, p. 89).

Adding to Harman's account, Lipton (2004) characterizes abduction as Inference to the Loveliest Potential Explanation (ILPE). The author explains that "we do not infer the best actual explanation; rather we infer that the best of the available potential explanations is an actual explanation" (Lipton, 2004, p. 58). There are two ways that a hypothesis can be considered the best potential explanation for a given evidence: the most warranted explanation, i.e. the *likeliest* to be true or most probable

explanation (Lipton, 2004, p. 59); and the *loveliest* explanation, which, if correct, would yield the deepest understanding (ibid.). In this proposal, Lipton offers a two-filtered characterization of ILPE: it generates new plausible hypotheses that *if true would explain the phenomenon*; and it selects among the available hypotheses the one that is the likeliest to provide the best explanation.

Lipton stresses that *explanatory considerations* constitute a guide for the generation of candidate hypotheses, as well as for the selection of the hypothesis that can produce the deepest potential understanding of the evidence to be explained (Lipton, 2004, p. 59). He (Lipton, 1999, p. 57, our highlight) stresses that "it is not simply that the phenomena to be explained provide reasons for inferring the explanations: we infer the explanations precisely because they *would*, if true, explain the phenomena". Campos nicely summarizes Lipton's account as follows:

> Likeliness is a measure – quantitative or qualitative – of the degree to which a general hypothesis agrees with all the evidence, so it is a measure of the inductive probability of the hypothesis. However, Lipton mainly advocates a model of inference to the *loveliest* potential explanation, that is, to the explanation that provides the deepest understanding of the phenomenon, and this is an abductive inference. (Campos, 2011, p. 440).

By focusing on the hypothesis' potential to provide the deepest understanding for a fact in need of an explanation, in terms of hypothesis generation and selection, Lipton's loveliness seems to come closer to the Peircean notion of abduction (cf. Paavola 2006, p. 98, Campos 2011). The main difference between the two approaches is that Peirce's abduction does not consider the potential likelihood of an abduced hypothesis; this would be designated in the pragmatic evaluation of the intrinsic value of a hypothesis within what Peirce (CP 6.528) calls "the economy of research". In contrast, Lipton's (2004, p. 71) approach aims to investigate "how loveliness helps to determine likeliness", i.e., how explanatory considerations should be considered as a criterion for the adoption of scientific hypotheses.

From Harman's and Lipton's approaches to IBE, the discussion of abduction and possible criteria for hypotheses' selection have received considerable attention in philosophy of science. Douven (2022, p. 44) summarizes IBE, stating that "[t]he core idea of abduction is often said to be that explanatory considerations have confirmation theoretic import, or that explanatory success is a (possibly fallible) mark of truth, or something similar". In other words, the conjecture of the truth of the hypothesis is *conditional* to its potential to explain the evidence in need of an explanation. Schurz (2023, p. 182) proposes a general pattern of abduction, as follows:

Premise 1: A singular or general fact E that is in need of explanation.
Premise 2: A system S of background beliefs, which implies that a certain hypothesis H is a (most) plausible potential explanation for E available in S ("potential" in the sense that if H were true, it would explain E).
Conclusion: H is conjectured to be true or at least close to the truth.

Some concerns emerge from the above characterization of abduction as IBE, such as the argument of the bad lot, used to inquire about the explanatory value in asserting the truth of a hypothesis that might be taken from a set of bad hypotheses (van Fraassen, 1989, p. 143). According to the bad lot argument, one cannot be epistemically justified to adopt a hypothesis as true simply in comparison to a set of available hypotheses. This is because the 'best' hypothesis can be chosen from a set of ill-defined hypotheses whose explanatory power is not satisfactory. In other words, the best explanation available could also be a bad candidate hypothesis. To solve this problem, one could argue that in IBE, one might consider the selected hypothesis as *good enough* to explain a given evidence, in relation to the available candidates (Dellsén, 2021). Adding to that, Bird (2010, p. 346) suggests that the selected

hypothesis should be significantly better in explaining E, among the available candidates. Bird (2017, p. 97, our highlights) stresses that "when we evaluate a scientific hypothesis, we may find that the evidence supports the hypothesis to *some degree*, but not to such degree that we would be willing to assert outright that the hypothesis is true".

To decide which hypothesis, among a set of plausible explanatory hypotheses, can be considered the best requires an account of explanatory virtues, such as simplicity, fruitfulness, scope, unification, and coherence, among others. The main problem is that the establishment of epistemic virtue might depend on the context of the fact to be explained in a given search domain.

In this context, we come back to our initial question: can abduction, understood as inference to the best explanation, be automated? We understand that if one has a well-structured set of criteria to represent a collection of epistemic virtues, for the establishment of what can be considered the best or the most probable explanation in a given context, we might have instances of automated forms of abduction. However, if the adoption of an explanatory hypothesis is considered highly context-dependent, attempts to automate abduction as IBE might depend on pre-established constraints that are built upon human decision-making processes and epistemic judgements. Thus, abduction as IBE could be seen as a *semi-automated* process of hypotheses selection. One might argue, however, that current algorithmic architectures are able to detect context-dependency and, thus, this is not a good argument against the impossibility to automate abductive reasoning. We understand, however, that the use of algorithmic models requires processes of data curation, analysis, and interpretation that, even if one could automate processes of generating and adopting explanatory hypotheses (see discussion in section 5), the initiative, curiosity and/or the attribution of relevance of a given fact to be explained, along with the complex set of parameters to be adjusted before running a model, depend on human strategies to engage in scientific inquiry. Thus, if we consider that, to some extent, IBE *requires* Peircean abductive reasoning, the same criteria for the impossibility of automation hold.

In the following, we explore the notion of abduction in relation to Bayesian Epistemology.

4 Abduction and bayesianism

Recent approaches to abduction center IBE in the context of Bayesian Epistemology (Bird, 2017; Feldbacher-Escamilla; Gebharter, 2019; Niiniluoto, 2022; Douven, 2022). The so-called compatibilist characterizations of abduction aim to discuss the extent to which an account of IBE would benefit from Bayesianism. The general idea is to combine IBE, which offers a descriptive account of abduction, with Bayesian Epistemology that aims to propose normative criteria for good inferential practices. According to Bird, compatibilism is:

> […] the view that people, scientists included, often hold the explanatory character of a hypothesis to be relevant to its epistemic evaluation. Inference to the Best Explanation (IBE), understood as a description of an inferential practice central to science, is explanationism *par excellence*: it holds that such subjects come to accept a hypothesis because it provides a better explanation of the evidence than its rivals. Per se, normative and descriptive claims are easy to reconcile. Even if IBE and Bayesianism are entirely different, it might be that IBE describes how we do reason while Bayesianism describes how we ought to reason (but do not). (Bird, 2017, p. 98).

In general, Bayesian Epistemology proposes that our beliefs come in degrees, meaning that the confidence one has in a given hypothesis can be measured in terms of probabilities. Bayesianism allows establishment of normative criteria for the attribution of probabilistic value to account for the degree of certainty of explanatory hypotheses. According to Bayesianism, rational agents ought to update their

degrees of belief in a given proposition in the face of new evidence, in accordance with Bayes' Theorem (Cabrera, 2017; 2022).

In an ideal scenario, a true belief receives a probabilistic weight equal to 1, while a false belief will receive a value of 0. The first rule of Bayesianism is that the probabilistic value attributed to the credence, or degree of belief, should be non-negative and the total sum equal to one. It undermines the chance of incurring Dutch Book arguments, i.e., "[…] a set of bets that are individually acceptable but jointly inflict a sure loss." (Lin, 2023, n.p.).

The second rule states that in facing new evidence, one ought to update one's degree of belief according to the principle of conditionalization, based on Bayes' rule that is formalized as follows: P(*h*|*e*) = P(*e*|*h*)P(*h*)/P(*e*). **P(*h*|*e*)** is called the posterior probability, i.e., the new probability attributed to the hypothesis *h* based on new evidence *e*. **P(*e*|*h*)** is the likelihood of the evidence *e* given hypothesis *h*. **P(*h*)** is prior probability of the hypothesis *h*, before evidence *e*. **P(*e*)** is the expectedness of the evidence, the prediction of *e* to happen. This is a normalizing constant that can be achieved by P(*e*|*h*)P(*h*) + P(*e*|*not-h*)P(*not-h*).[2]

The investigation of IBE in relation to Bayesianism helps the search for rational grounds for explanatory inference, going beyond the descriptive approach of IBE: it enables to consider not simply the most probable hypothesis given new evidence, but the most probable hypothesis in relation to the most fruitful ones. For example, Douven (2022) delves into a Bayesian approach to abduction, where he adds to the Bayesian conditionalization norm a criterion *c* to account for explanatory goodness of a hypothesis. Douven explains that:

> This means that actually they can assign bonus points as well as malus points; where a hypothesis is an extremely poor explanation of the evidence, they can even assign a malus point of -I, which when add to the hypothesis's probability could result in negative value unsuitable for "normalizing" to a probability. (Douven, 2022, p, 169).

This account of abduction requires a given set of hypotheses for the attribution of explanatory goodness, instead of the attribution of probabilistic values to each hypothesis separately. The author proposes that explanatory goodness should be measured within a *range* from -1 to 1, with 0 being the neutral value. However, what counts as objective criteria for explanatory goodness must be determined. For example, Douven (2022, p. 89) applies Popper's and Good's measure of explanatory goodness as the criterion (see also Douven; Schupbach, 2015 for details).

Another example of Compatibilism can be found in Lipton (2004) and in Bird's (2017) interpretation of Lipton's suggestion. The authors agree with the notion that the Bayes principle of conditionalization provides a probabilistic account for the evolution of the degrees of belief one has in a hypothesis, given evidence. However, the attribution of values that makes feasible the conditionalization of one's beliefs is not always clear. Lipton (2004) suggests accounting for explanatory considerations in the transition from prior to posterior assessments. Informed by his account of IBE, the author suggests that the determination of the likelihood **P(*e*|*h*)** should consider the explanatory loveliness of *h* in explaining evidence *e*. Subsequently, by conditionalization, the likelihood, informed by explanatory considerations, would guide likeliness or the posterior probability (remember that, for Lipton, loveliness is a guide for likeliness). The author explains:

> The present proposal is that the mechanism by which this works may be understood in part by seeing the process as operating in two stages. Explanatory loveliness is used as a symptom of the likelihood (the probability of E given H), and the likelihoods help to determine likeliness or posterior probability. (Lipton, 2004, p. 115).

2 The authors would like to thank Cassie Bird and Francisco Camargo for helping with the mathematical technicalities.

According to Lipton's (2004) and Bird's (2017) compatibilism, one can also devise *heuristic rules* for prior attribution to account for the explanatory potential of a hypothesis before conditionalization. In this account, one should assess the explanatory loveliness of the prior value of a hypothesis P(*h*), as well as the expectedness of the evidence P(*e*). For example, considering features such as unification, simplicity, scope, and other explanatory virtues. Subsequently, one should evaluate the explanatory loveliness of *h* in relation to *e*, and change the explanatory assessment of *h* accordingly. This step corresponds to the conditionalization principle based on Bayes' rule.

Finally, as a form of posterior assessment, one can determine which evidence is relevant to the hypothesis to be considered in the conditionalization process. Bird (2017, p. 100, our highlights) explains that "the injunction to consider the total evidence, although implicit in Bayesianism, is not one we can actually implement, so we need a heuristic to guide us to the relevant evidence, viz. whether the evidence could be explained by the hypothesis". Thus, this approach suggests three roles of explanatory considerations for Bayesian conditionalization: the determination of the likelihood, the consideration of prior probability of the hypothesis and the evidence, and the determination of relevant evidence (Lipton, 2004, p. 114).

In summary, compatibilism advocated to abduction combines probabilistic measures and explanatory goodness in the establishment of the degree of belief one attributes to a given hypothesis given new evidence. So, can abductive inference be automated based on Bayesianism? If we take compatibilist accounts of abduction, we end up with the same problem we face in IBE: to define which epistemic virtues constitute a good representation of what can be considered the "best" hypothesis in a given context. However, if we embrace the Bayes rule as a form of updating belief, and we call it abduction, there would be no difficulty with the possibility that "abduction" could be automated, given that the Bayes rule is a probabilistic formula. The problem of assuming Bayes updating rules as abductive reasoning is that it conflates abduction with induction: Bayes' rules do not allow the generation or adoption of an explanatory hypothesis, but simply the determination of a value based on prior probabilities. It is the attribution of a hypothesized value as a prior probability that could be seen to require Peircean abduction.

In the following, we offer an example of an attempt to automate scientific discovery by considering the possibility of taking abduction as a logical module within the software architecture.

5 Automating discovery: an example from neuro-symbolic AI

*We believe that AI-Descartes is a promising step towards achieving the ultimate goal of understanding and explaining the world.*3

In 2023, IBM launched AI-Descartes, an AI system that combines logical reasoning with symbolic regression, aiming to derive scientific discovery from axiomatic knowledge and experimental data (Cornelio et al., 2023). As one can note in the above quotation, the authors seem very optimistic about the future of their framework for automating scientific discovery. Their efforts are part of the developments in neuro-symbolic AI that can be generally characterized as hybrid systems that combine statistical algorithms and logical inferences.[4] The idea underlying AI-Descartes is to find latent patterns in massive amounts of data, along with constraints that will trim the relevant findings by means of logical reasoning. As part of the logical module of the system's architecture, the authors consider abductive inference as a form of logical technique for hypothesis generating. However, they do not further develop the idea in the paper. In the following, we present the main components of AI-Descartes

3 https://research.ibm.com/blog/ai-descartes-scientific-discovery.

4 However popular the developments on neuro-symbolic AI are becoming, its meaning is still ambiguous (Submann et al., 2023, p. 12).

to investigate what could be the role of abduction in their framework, also inquiring about the extent to which abduction could be automated.

In general terms, the system is designed to automate the discovery of an *unknown symbolic model* (i.e. a formula) that fits a collection of real data points (extracted from existent datasets), being also derivable from background theory (a set of pre-established axioms). The aim is to "obtain hypotheses from data and assess them against theory" instead of "obtain hypotheses from theory and then check them against data" (Cornelio et al., 2023, p. 3). In this framework, a "hypothesis" means a symbolic model or candidate formula, and "explanation" means a missing axiom in an incomplete background theory.

As described in Cornelio et al. (2023), AI-Descartes has four main elements: (i) *Background knowledge*, which comprises a set of domain-specific axioms. This set is expected to be *logically complete* (by encompassing the axioms necessary for the explanation of the suggested formula) and *consistent* (the axioms do not contradict each other). (ii) *A class of hypotheses* composed by symbolic models and logic axioms that are defined by a grammar and constraints (e.g., to avoid redundancy and guarantee monotonicity). (iii) *Data* that represent a set of examples including the values of dependent and independent variables (expressing y as a function of x: $y = f^*x$). (iv) *Modeler preferences*, described as a set of parameters (such as an error function, accuracy, complexity).[5]

The system is designed to find, from a dataset and background knowledge, possible candidate symbolic models by means of an optimized form of symbolic regression. The optimization process occurs in a higher level where the programmer chooses different parameters to implement in the symbolic regression, such as the structure of a generalizable expression tree, its length and depth, number of branches, the numerical function, set of invariants, and so on (Cornelio et al., 2023, supplementary information). After choosing the parameters, an *optimization problem* is settled and the program is ready to be run by an automated logical solver (in AI-Descartes, the authors opted to apply an *MINLP*[6]*-Based Symbolic Regression solver*, with the logical reasoning systems *KeYmaera X* and *Mathematica* that are automated theorem provers (Cornelio et al., 2023, p. 3)).

Within this process, the program is able to come up with a set of candidate formulae that potentially fit the data (ibid, p. 3). Then, for each candidate formula, the system calculates a distance function (or error) to measure the extent to which the symbolic formula is derivable from background knowledge. In other words, it measures the distance between the candidate model *f* that would fit the data *and* the extent to which it can be derivable from background knowledge *fb*. From this, the system gives as output the chosen formula (or model) with its distance error or a proof of inconsistency (i.e., that although the formula fits the data, it is not derivable from background knowledge). If there are no derivable candidates, the system might require additional data, the revision of the adopted constraints, changes in the background theory, or the generation of candidate axioms to be fed into the background theory. With this framework, the authors claim that their system "[…] yields an end-to-end discovery system, which extracts formulas from SR [symbolic regression], and furnishes either a formal proof of derivability of the formula from a set of axioms, or a proof of inconsistency" (ibid, p. 2).

In a nutshell, the system proposed by Cornelio et al. (2023) is developed to find a formula from the data and then check the list of potential generated hypotheses against the derivable properties of the background theory. Error measurement techniques are developed to measure the compatibility of the generated symbolic model (formula/hypothesis) against the data and the background knowledge. If candidate models fit the data but are not derivable from the background knowledge, the abductive module will be designed to generate new axioms that will be fed into the background knowledge and the process will iterate again. Although the authors have included abduction as part of the reasoning module of AI-Descartes, they did not implement abduction in the current version of the system.

---

5 A detailed explanation of the software functioning can be found at: <https://www.youtube.com/watch?v=oIzq8iAO6wA>

6 MINLP stands for mixed-integer nonlinear programming.

Abduction would be ideally employed in the above framework to account for cases when the background knowledge is incomplete, i.e., when a given discovered formula cannot be derivable from the available information. The authors describe abduction as a technique to find explanations, understood as missing axioms, given a logical theory. The generated explanation will, in turn, enhance the incomplete background theory. The authors emphasize that "the explanation axioms are produced in a way that satisfy the following: (1) the explanation axioms are consistent with the original logical theory and (2) the observation [the formulae extracted from numerical data] can be deduced by the new enhanced theory (the original logical theory combined with the explanation axioms)" (Cornelio et al., 2023, p. 7).

The authors validate their framework by testing it in three discoveries from physics, one of which was to derive Kepler's third law of planetary motion, which describes the orbits of the planets around the sun. From three real world databases, and with the Newtonian law of motions as background theory, AI-Descartes developers claim to rediscover the equation of Kepler's third law. With the datasets and the input of a set of operators in the SR system, they arrived at a list of possible hypotheses that corresponded to approximated candidate formulae, evaluated by the development of error measurement techniques that calculate the distance between the suggested hypotheses and the derivable formulae.

We consider that the system has at least three limitations. The first one is the underdevelopment of the *optimal experimental design* module that aims to account for the relevant experiments that might contribute to the process of discovery in a given domain. Although the authors do not implement both abductive module and experimental design in their initial framework, we believe that the prospect of such implementations, along with a common-sense logic module, looks likely to occur. The second limitation is regarding the *data format* for the input. Only numerical data is being considered, which constrains the system's subject-area of application. Another limitation worth considering concerns the *disciplinary scope* that forms the background knowledge. The examples given are from physics, but how this framework would perform in other disciplines with assumptions that are not axiomatized is something to be investigated.

Given this preliminary description of AI-Descartes, we can come back to our question: can abductive inferences be automated? As one could expect, human judgment permeates the decision of parameters, background theory, the forms for calculation of the distance errors, and the questions to be asked to the system. The developers of AI-Descartes understand abduction within the framework of Logic Programming.[7] Denecker and Kakas (2002, p. 404) stress that a common characterization of abduction in formal logic is described as: "Given a logical theory *T* representing the expert knowledge and a formula *Q* representing an observation on the problem domain, abductive inference searches for an explanation formula *E*". In other words, an abductive logic is designed to find an explanation *E* for the observed fact to be explained *Q* within theory *T*. Here, abductive logic is taken to be a Harmanian instance of IBE and, as we have discussed in Section 3, it can only be semi-automated, as it depends on the decision of the parameters and constraints to be implemented in the system.

Although the abduction module is missing in the current state of the summarized system, one could say that in its current form, AI-Descartes might perform a partial form of abduction, as it is able to generate hypotheses from the data, and select them, based on reasoning errors and derivability. On the one hand, AI-Descartes indeed generates and selects hypotheses based on data. On the other hand, it does so mechanically: there is no reasonability in choosing a hypothesis. So, it could be said to perform inference to the best explanation, but not abduction in the Peircean sense, which requires the recognition of a problem in a surprising situation and the establishment of relevance in the adoption of a hypothesis (whatever relevance means). The tricky aspect of attributing relevance in the adoption of an explanatory hypothesis has to do with the contextual aspect of scientific reasoning. "What is a context?" is a hard question to answer, and even harder to be computationally implemented.

7 Personal communication.

6 Concluding provisional remarks

In this paper, the possibility of automating explanatory inferences has been investigated in the context of scientific discovery, with focus on distinct accounts of abductive reasoning. As an example, we introduce the AI-Descartes framework that seemingly implements abductive reasoning as part of its architecture. As we have stressed, abduction is involved in processes of discovery, creativity, problem solving, question-answering, insight, hypothesis selection, and explanation, among others. We also indicated that according to Peirce, abductive inference initiates with the feeling of surprise, and it is developed with the search for explanatory hypotheses that, if recognized as a reasonable support to the fact in search for an explanation, would dissipate this type of feeling. Simulations of the feeling of surprise by actors can be successful in specific contexts, such as in the theater, cinema, and even in real life, but they all require a context to make sense. It is not clear (yet?) in which circumstances this feeling could be implemented in a machine.

We argued, however, that the generation of candidate hypotheses to dissipate our feeling of surprise could be partially automated by the implementation of AI tools. If we - as humans - do not have the capacity to come up with trillions and trillions of possible explanatory hypotheses in a short time span, to render a phenomenon unsurprising, we can create mathematical constructs that might help us with this task. In this sense, computational tools could enhance our cognitive abilities to formulate and select potential explanatory hypotheses in the process of abduction.

The question to pose would be: to what extent might automation change the human ability to perform abduction? In other words, what are the implications of the growing automation of scientific reasoning for the scientists' cognition? Is the capacity to experience spontaneous surprise in real life worthy of care and admiration? Perhaps, more considerations on these questions are what we can learn from the study of abductive reasoning.

Acknowledgements

We would like to thank FAPESP for support of the present research (project numbers 2020/03134-1 and 2023/01405-8). We thank the two anonymous reviewers for their valuable comments and suggestions. We would also like to thank the members of the Academic Group of Cognitive Studies (GAEC) and the members of the Egenis Research Exchange for the valuable contributions and carefully reading and commenting on the earlier version of this paper.

References

ANDERSON, D.R. The Evolution of Peirce's Concept of Abduction. *Transactions of the Charles S. Peirce Society*, v. 22, n. 2, p. 145-164, 1986. http://www.jstor.org/stable/40320131

BAIN, A. *Mental and Moral Science*, 3rd edn. London: Longmans, Green and Co., 1872. Bk 4, Ch. 8, p. 371-385.

BELLUCCI, F. Eco and Peirce on Abduction. *European Journal of Pragmatism and American Philosophy*, 2018. https://doi.org/10.4000/ejpap.1122.

BELLUCCI, F.; PIETARINEN, A.-V. Icons, Interrogations, and Graphs: On Peirce's Integrated Notion of Abduction. *Transactions of the Charles S. Peirce Society*, v. 56, n. 1, p. 43-61, 2020. https://doi.org/10.2979/trancharpeirsoc.56.1.03

BIRD, A. Eliminative Abduction: examples from medicine. In: Studies in History and Philosophy of Science, Vol. 41, 2010. 345–352. https://doi.org/10.1016/j.shpsa.2010.10.009

BIRD, A. Inference to the Best Explanation, Bayesianism, and Knowledge. *In:* MCCAIN; POSTON (Eds.). *Best Explanations*: New Essays on Inference to the Best Explanation, 2017. https://doi.org/10.1093/oso/9780198746904.003.0007, accessed 31 Oct. 2023.

CABRERA, F. Can there be a Bayesian explanationism? On the prospects of a productive partnership. *Synthese* v. 194, p. 1245–1272, 2017. https://doi.org/10.1007/s11229-015-0990-z

CABRERA, F. Inference to the Best Explanation: an Overview. *In*: MAGNANI, L. (ed.). *Handbook of Abductive*

Cognition. Springer, Cham., 2022. https://doi.org/10.1007/978-3-030-68436-5_77-1

CAMPOS, D.G. Imagination, Concentration, and Generalization: Peirce on the Reasoning Abilities of the Mathematician. *Transactions of the Charles S. Peirce Society*, v. 45, n. 2, p. 135-156, 2009. https://doi.org/10.2979/tra.2009.45.2.135.

CAMPOS, D. On the distinction between Peirce's abduction and Lipton's inference to the best explanation. *Synthese*, v. 180, p. 419-442, 2011. https://doi.org/10.1007/s11229-009-9709-3

CORNELIO, C. et al. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nat Commun*, v. 14, p. 1777, 2023. https://doi.org/10.1038/s41467-023-37236-y

DELLSÉN, F. Explanatory Consolidation: from best to good enough. *Philosophy and Phenomenological research*, v. 103, p. 157-177, 2021. https://doi.org/10.1111/phpr.12706

DENECKER, M.; KAKAS, A. Abduction in Logic Programming. *In*: KAKAS, A.C., SADRI, F. (Eds.). *Computational Logic*: Logic Programming and Beyond. Lecture Notes in Computer Science, vol 2407. Springer, Berlin, Heidelberg, 2002. https://doi.org/10.1007/3-540-45628-7_16

DOUVEN, I.; SCHUPBACH, J. N. Probabilistic alternatives to Bayesianism: the case of explanationism. Frontiers in Psychology, p. 1-9, 2015. https://doi.org/10.3389/fpsyg.2015.00459

DOUVEN, I. *The art of abduction*. MIT Press Direct, 2022. https://doi.org/10.7551/mitpress/14179.001.0001

FELDBACHER-ESCAMILLA, C.J.; GEBHARTER, A. Modeling creative abduction Bayesian style. *Euro Jnl Phil Sci,* v. 9, 2019. https://doi.org/10.1007/s13194-018-0234-4

GROOVER, M.P. automation. *Encyclopedia Britannica*, 22 Oct. 2020, <https://www.britannica.com/technology/automation. Accessed in 2 May 2022.

HARMAN, G.H. The Inference to the Best Explanation. *Philosophical Review*, v. 74, n. 1, p. 88-95, 1965. https://doi.org/10.2307/2183532

HINTIKKA, J. What Is Abduction? The Fundamental Problem of Contemporary Epistemology. *Transactions of the Charles S. Peirce Society*, v. 34, n. 3, 1998. http://www.jstor.org/stable/40320712

IBRI, I. The heuristic exclusivity of abduction in Peirce's philosophy. In: LEO, R. F.; MARIETTI, S. (Org.). *Semiotics and Philosophy in C. S. Peirce*. Cambridge: Cambridge Scholars, 2006. p. 89-111.

JOSEPHSON, J.; JOSEPHSON, S. *Abductive Inference*. Cambridge University Press, 1994.

LIN, H. Bayesian Epistemology. *In*: ZALTA, E. N.; NODELMAN, U. (Eds.). *The Stanford Encyclopedia of Philosophy.* Winter, 2023.URL = <https://plato.stanford.edu/archives/win2023/entries/epistemology-bayesian/>.

LINDSAY, R. K. et al. DENTRAL: a case study of the first expert system for scientific hypothesis formation. *In*: Artificial Intelligence. Elsevier, Vol. 61, 1993. p. 209-261,

LIPTON, P. *Inference to the Best Explanation*. New York. Routledge, 1999.

LIPTON, P. *Inference to the Best Explanation*. 2 ed. Edition: London; New York. Routledge, 2004.

MINNAMEIER. G. Forms of Abduction and an Inferential Taxonomy. *In*: MAGNANI; BERTOLOTTI (Eds.). Handbook of Model-Based Science. Springer, 2017. p. 175-195.

NIINILUOTO, I. Explicating Inference to the Best Explanation. *In*: GONZALEZ, W. J. (Ed.). *Current Trends in Philosophy of Science*. Synthese Library, vol 462. Springer, Cham., 2022 https://doi.org/10.1007/978-3-031-01315-7_11

PAAVOLA, S. Hansonian and Harmanian Abduction as Models of Discovery. *International Studies in the Philosophy of Science*, v. 20, p. 93-108, 2006. https://doi.org/10.1080/02698590600641065

PAAVOLA, S. *On the origin of ideas*: an abductivist approach to discovery. Revised and enlarged edition. Saarbrücken: Lap Lambert Academic Publishing, 2012.

PEIRCE, C.S. The Collected Papers of Charles Sanders Peirce. Electronic edition. Vols. I-VI, HARTSHORNE, C., WEISS, P. (Eds.), 1931-1935. Vols. VII-VIII, Burks, A. W. (Ed.). Charlottesville: Intelex Corporation. Cambridge: Harvard University Press, 1958. [Quoted as CP, followed by the volume and paragraph].

PEIRCE, C.S. The New Elements of Mathematics by Charles S. Peirce, Vol. 4., Eisele, C (ed.). Mouton, The Hague, Paris, 1976. [Cited as NEM followed by volume and page number].

PEIRCE, C.S. The Essential Peirce: Selected Philosophical Writings. Vol. 2 (1893-1913). Peirce Edition Project (Eds.). Bloomington & Indianapolis: Indiana University Press, 1998.

PEIRCE, C.S. Writings of Charles S. Peirce: A Chronological Edition: 1886-1890, Vol. 6, Houser, N. et al. (eds). Indiana University Press: Bloomington & Indianapolis, 2000.

PIETARINEN, A.V.; BELLUCCI, F. The Iconic Moment. Towards a Peircean Theory of Diagrammatic Imagination. *In*: REDMOND J., MARTINS, O. P., FERNÁNDEZ, Á. N. (Eds.). *Epistemology, Knowledge and the Impact of Interaction*. Logic, Epistemology, and the Unity of Science, vol 38. Springer, Cham., 2016. https://doi.org/10.1007/978-3-319-26506-3_21

SANTAELLA, L. Abduction: the logic of guessing. *Semiotica,* v. 153, p. 175-198. 2005. https://doi.org/10.1515/semi.2005.2005.153-1-4.175

SCHURZ, G. Theory-Generating Abduction and Its Justification. In: *Handbook of Abductive Cognition*, Magnani, L. (ed.). Springer, Cham, 2023, p. 181–208. https://doi.org/10.1007/978-3-031-10135-9_4

STJERNFELT, F. *Diagrammatology*: An investigation on the borderlines of phenomenology, ontology, and semiotics. New York: Springer, 2007.

STJERNFELT, F. Three Tacit Gossipers: A Few Symbol Strings Regarding New ai and Old Philosophy. *Danish Yearbook of Philosophy*, v. 57, p. 100-115. 2024. https://doi.org/10.1163/24689300-bja10054

SUDMANN, A. et al. Research with Subsymbolic AI. In: Beyond Quantity: research with subsymbolic AI. Sudmann, A. et al. (eds.), Bielefeld: Majuskel Medienproduktion GmbH, Wetzlar. AI Critique. 2023, p. 33-60,

VAN FRAASSEN, B. C. *Laws and Symmetries*. Oxford: Oxford University Press, 1989.

VEALE, T.; CARDOSO, A.; PEREZ, R.P. Systematizing Creativity: a computational view. *In*: VEALE & CARDOSO (Eds.). *Computational creativity*: the philosophy and engineering of autonomously creative systems. chapter 1. Springer Nature Switzerland, 2019. https://doi.org/10.1007/978-3-319-43610-4

OPEN ACCESS

COGNITIO

Revista de Filosofia
Centro de Estudos de Pragmatismo

São Paulo, v. 25, n. 1, p. 1-15, jan.-dez. 2024
e-ISSN: 2316-5278

https://doi.org/10.23925/2316-5278.2021v22i1:e68263