Riegler, A. (in press) Is a closed-loop discovery system feasible? In Magnani , L. (ed.) *Proceedings of E-CAP 2004*.

Is a Closed-Loop Discovery System Feasible?

Alexander Riegler
CLEA, VUB
Krijgskundestraat 33, B-1160 Brussels, Belgium
ariegler@vub.ac.be

Abstract. In order to construct scientifically reasoning artifacts we not only have to close the loop between hypothesis generation and evaluation but also to make the system embodied. To genuinely understand scientific insights, "robot scientists" need to represent scientific knowledge within their own representational structure rather than in terms of a priori defined logical propositions. Two main features of such systems are identified: projective constructivism that reverses the flow of information processing, and cognitive canalization that reduces computational requirements.

Key words. Cognitive canalization, embodiment, projective constructivism, representation, robot scientist, white noise.

Introduction

Recently, philosophy of science has turned towards the cognitive aspects of scientific activity, i.e., it has started to focus on the human scientist who is carrying out science (e.g., Giere 1993; Carruthers, Stich & Siegal 2002). The basic assumption of this program is that the same general cognitive processes serve as a vehicle for both scientific and non-scientific thinking. Not only can such a psychology of science account for the creative aspects of science—as opposed to a merely rationalist-logical system in which according to its definition as being a deductive system no creativity is possible—it also enables us to think about another perspective: Can machines perform creative science as well? Can a psychology of science provide insights and mechanisms that—in the long run—can be automatized and therefore passed over to computational devices in order to carry out scientific reasoning?

In this paper, I will explore the road to scientifically reasoning artifacts based on insights from cognitive science and epistemology. These artifacts are supposed to go beyond the current application of computers that are used as a supportive tool in virtually all disciplines, whether as "number crunchers" in mathematics and physics, or as databases to store large amounts of data. In both cases they are used to cover deductive facets of scientific activity. A typical example is the proof of the famed four-color conjecture (Appel & Haken 1977), which demonstrated that using supercomputers to calculate individual cases of the problem is an alternative to proving the problem in a traditional mathematical way. In contrast to humans, however, this program did not come up with the four-color conjecture in the first place. It merely tracked it unremittingly down.

More sophisticated discovery systems have been around for many years already, such as the program "Bacon" (Langley et al. 1987). I will argue that these systems, while being a first important step, are flawed in the sense that they are programs whose input is fed by humans and whose computational output is interpreted by humans. For scientifically reasoning machines,

however, creativity does not consist of simply re-arranging existing *anthropomorphic* pieces in novel ways.

In order to arrive at a more promising approach, I will discuss the problems of present-day discovery systems, including the recent "robot scientist" of King et al. (2004). Building on my arguments, I will introduce two principles that can be considered basic ingredients of discovery systems: projective constructivism and cognitive canalization. Finally, I will provide a synopsis of an algorithm based on these principles.

Problems of discovery programs

Data-driven discovery programs have been used in various scientific domains. Many of these systems perform equation discovery of quantitative laws and have successfully reproduced historical findings.

However, there are severe limitations to such empirically working systems. They face the problem that there is a *practically infinite* amount of ways to extract laws from a given data set, known as the empirical underdeterminism of theories. Systems such as "Bacon" (Langley et al. 1987) have circumvented the problem by using appropriately pre-prepared data. However, this procedure does not comply with historical discoveries made by humans where part of the problem was to first find the *relevant data* in the bulk of measured data. Johannes Kepler, for example, had first to wade through a huge amount of data collected by his predecessor Tycho Brahe and himself over many years (Kozhamthadam 1994). Then he had to make the "right" choice, namely to tell relevant data from irrelevant, in order to find the geometrical figure that represents the orbits of the planets. It took him *thirteen* years to come up with the idea that the data he had about the movements of the planet Mars fit with the concept of planets revolving around the sun in ellipses. Kepler's case gives rise to the suspicion that pruning a deluge of data might not be a rational process at all and thus difficult to implement in machines. Furthermore, from the implementation point of view, there is the problem of how to represent empirical observations, as in any real-world environment the number of possible propositions and their mutual relationships are practically infinite (the "frame-problem," e.g., Dennett 1984).

The fact that human science has reached high standards despite the severe limitations of human cognition (small short-term memory, emotionally biased evaluation, mental inertia, etc.; e.g., Riegler 1998) gives rise to the assumption that human scientists—like human chess players (Chase & Simon 1973)—do not use brute "computational" force in order to arrive at new discoveries. The success of human science can be partly explained by the social dynamics of consensus-seeking and controversy within scientific communities. The challenge, however, is to create *machines* that eventually surpass the limitations of human scientific reasoning.

The dependency of present-day computer programs on humans regarding data preparation and interpretation contributes little to this challenge. Therefore, it has been proposed to "close the loop" and let the program not only generate and select hypotheses, but also carry out the necessary experiments (Hayes-Roth 1983; Bryant et al. 1999). Very recently, King et al. (2004) claimed to have successfully implemented a closed-loop system. They developed a "robot scientist," a system that "automatically originates hypotheses to explain observations, devises experiments to test these hypotheses, physically runs the experiments using a laboratory robot, interprets the results to falsify hypotheses inconsistent with the data, and then repeats the cycle." The specific goal of the robot scientist—to determine the function of genes from the performance of knockout mutants—was implemented as the interplay of abduction and deduction; the former infers missing chemical reactions that could explain observed phenotypes, the latter checks the consistency of hypotheses generated. The authors emphasize that the "key point is that there was no human intellectual input in the design of experiments or the interpretation of data." However, a close look at how the system represents *prior* biological

knowledge reveals a rather anthropomorphic picture: nodes in a directed graph denote metabolites and its arcs are interpreted as enzymes. This results in a set of logical propositions, which are used by the programming language Prolog to compute predictions. So while the actual processing cycle might indeed be "human-free," the algorithm itself is disembodied—a criticism that has been applied to symbolic artificial intelligence for many decades. The problem with disembodied propositions is a two-fold arbitrariness: (1) the arbitrariness of grounding propositions in the world (Harnad 1990); and (2) the arbitrariness of linking up propositions in a sensible way, as the following example illustrates. Assume that in an astronomical discovery system a proposition such as P1: "perceive (Mars, x, y)" leads to another proposition, say, P2: "move (telescope, x', y')." How can P2 be contingent on P1 or, paraphrasing Heinrich Hertz (1894/1994), how can P2 be a "*denknotwendige*" (logically necessary) consequence of P1? The transition appears arbitrary; it is as "semantically blind" as the scientific device in Jonathan Swift's *Gulliver's Travels, A Voyage To Laputa*. This scientific machine could make "a complete body of all arts and sciences" through the mechanical, i.e., syntactical, combination of all words in a language. The task of the human user was to pick out the meaningful sentences. Obviously, this sort of machines is not able to do that.

So what does it take to create closed-loop discovery systems that display semantic competence and therefore properly *represent* and *understand* scientific insights? Quite obvious syntactical reconstructions of phenomena are insufficient. Even the most sophisticated computer simulations are *anthropomorphically* designed by the programmer rather than having the degree of autonomy natural systems enjoy. For example, in artificial life programs, the simulated creatures react according to a priori specified rules rather than behavioral patterns, which are the result of phylogenetic and ontogenetic developments. Natural animals seem to "understand" their environment rather than to copy their behavioral repertoire from biological textbooks like artificial life programmers do. The crucial aspect is that understanding involves a form of coupling between agent and system rather than mechanical–algorithmic reproduction of some recipes (Riegler 2002).

Projective constructivism

Given all the obstacles discussed in the previous sections, Immanuel Kant's (1781/1991) "Copernican Turn" points in the direction of a possible solution. Kant suggested that "objects must conform to our knowledge" rather than the other way around, which considers knowledge a mirror of the state of affairs in the "objective world." Kant's idea radically dismisses any form of determinism of the cognizing individual through the outside reality (cf. also Bettoni 1997). In other words, it rejects the inductive-empirical mode of knowledge acquisition according to which an organism "extracts" information from an environment and passes this "raw" primary representation onto further "processing" within the cognitive apparatus. This is what contemporary image-processing and data-mining algorithms do: They check mechanically for regularities in complex patterns. Only through spatio-temporal embedding, or "structural coupling" (Maturana & Varela 1980), can a system capture essential features that allow for successful prediction (Riegler 2002). This holds true especially for cognition but also for scientific models. Consequently, I claim that neither mathematics (Wigner 1960) nor cognition (Wolff 1982; Chater & Vitány 2003) articulate a (compressed) description of "reality". Rather, "reality" is brought forward *in terms of* cognitive constructions such as (mathematical, computational, formal, etc.) structures. I would like to call this assertion "*projective constructivism*."

Intuitively, projective constructivism holds responsible for the fact that people have always perceived heroic figures and animals in certain stellar constellations. That is, they projected

some internally generated structure onto a pattern of lights in the night sky. Similarly, people may perceive faces in the random formations of clouds, etc.

Projective constructivism can be considered the opposite of John Searle's (1980) picture of artificial (and human) intelligent systems in his well-known Chinese room argument. The person in the room receives characters from outside the room, processes these characters according to a rulebook, and passes the re-written characters out of the room. Projective constructivism, however, emphasizes that the internal working of the cognitive system (i.e., the inmate's cognitive activity) comes first. It generates (mental) structures—the "rules"—in the first place, which are subsequently ascribed to the characters that arrive from outside. In other words, the internal "logics" is mapped onto what is perceived as the "outside world."

What are the consequences for scientifically reasoning artifacts? Instead of having artifacts extracting features from their environment and constructing some models upon which to base their reasoning, I propose to reverse the direction and have artifacts project a priori mental structures onto "external" sensorimotor experiences. This ties in well with the psychological insights of Ulric Neisser (1976). Neisser characterized perception as a schemata-controlled "information pickup," i.e., the organism's cognitive apparatus (schemata) construct anticipations of what to expect and thus enable the organism to actually perceive the expected information. What is not anticipated cannot be perceived. For example, a circle drawn in sand is perceived as a circle not because of sophisticated image processing in our head, which compresses the perceived trace into the mathematical concept of a circle, but due to the projection of a mathematically ideal circle onto sensory data.

Projective constructivism receives empirical support from experiments regarding "superstitious perception" (Gosselin & Schyns 2003). The authors stimulated the visual system of test subjects with unstructured white noise, i.e., a static bit pattern that has equal energy at all spatial frequencies and does not correlate across trials. The subjects were asked to discriminate between a smiling and a nonsmiling face, which was allegedly present in 50% of the presentations. As a result, the subjects perceived the expected face. These findings confirm projective constructivism in the sense that the anticipated pattern was projected onto (partially correlated with) the perception of the white noise.

Also in ethological experiments we find corroborating results. B. F. Skinner's 1948 article on "superstition in the pigeon" describes how birds react in situations beyond their cognitive control. Skinner presented food to hungry pigeons at regular intervals, with no reference whatsoever to their current behavior. Soon the birds started to display certain rituals between the reinforcements, such as turning two or three times about the cage, bobbing their head, and incomplete pecking movements. As Skinner remarked, the birds happened to be executing some response as the food appeared the first time, and they tended to repeat this response if the feeding interval was only short enough. In a certain sense, the pigeons projected the idea of a link between behavior and feeding onto their behavioral display.

Cognitive canalization

The apparently erroneous projections in these experiments beg the following question. Does not projective constructivism imply some degree of arbitrariness in the sense that the mind may construct anything it fancies? For obvious reasons, we are compelled to assume that the construction of mental structures that precedes actual perception and cognition must not be arbitrary; otherwise the mind would drown in a sea of solipsistic structurelessness. In the context of robot scientists, this would result in "solipsistic machines." However, as Piaget (e.g., 1954) and other developmental psychologists have pointed out, the cognitive faculty develops gradually over time rather than being instantiated at once. If we assume that cognition is based on constructions, these mental constructions must be regarded as historical assemblies. Their

historicity imposes a hierarchical structure among the components (Simon 1969) in which more recent additions attach to older (and preferentially bigger) ones. Such a hierarchy results in mutual dependencies among its components. Removing one component will not only change the context (i.e., configuration and connections) of other components but may even destroy the accessibility of those components altogether if the removed part served as a hub in the sense of Barabási (2002). The dependencies result in canalization. It severely restricts the degrees of freedom in the way future constructions can be added. If added in wrong places, they would cause disruption or even disintegration. In the context of cognitive systems, we can speak of self-generated *cognitive canalization*, which prevents the constructions of the mind from being arbitrary (Riegler 2001b).

Cognitive canalization can be compared with the effect an ever expanding jigsaw puzzle has on newly added pieces. Each piece that has found a place where it fits locks in with its neighbors. By doing so, it also expands the puzzle's border, which in turn enables the addition of further pieces. On the one hand, the more pieces lock in the larger the puzzle becomes and the more pieces can be attached. But on the other hand, the actual shape of the expanding border of the puzzle determines which sort of pieces can be attached next. Consequently, canalization results in irreversibility (or asymmetry). Systems, whether natural or artificial, are driven into a continuous complexification of their structure, thus yielding asymmetry in time, caused by internalist rather than externalist mechanisms (Riegler 2001a).

Synopsis of the algorithm

A prototype that implements both principles is the "constructivist–anticipatory algorithm" (cf. Riegler 1994). Lack of space prevents a detailed description of the algorithm but it can be best characterized as a semi-neuronal production system, which, in contrast to programs based on logical inferences over propositions, does not a priori specify how knowledge is represented. In order to avoid the fallacy of anthropomorphically defined logical propositions, the basic representational structure are schemata, i.e., compounds of conditions and sequences of actions, working on memory cells. In the spirit of Mach's (1897) and Bridgman's (1927) operationalism, schemata represent sensorimotor knowledge, i.e., the anticipation of how to handle things under certain conditions with the proper action sequence. The condition part provides context matching which allows the schema that best fits the present context to execute its action sequence. As soon as a schema finishes, context matching starts again.

The aspect of projective constructivism is implemented as the evaluation of schemata. Once invoked, the schemata ask for sensory or internal data only when they need them. In other words, the algorithm, which projects its dynamical structure outward, neglects environmental events except for the demands of the current action sequence. This leads to a significant decrease in computational costs, since the agent equipped with the algorithm need not extract the full environmental information each time step. The algorithm is in sharp contrast to the information-processing paradigm that defines the cognitive system as a bottleneck: the essential features would need to be selected among the wealth of "information" provided by the "outside" in order to decrease the enormous degree of complexity.

However, the system goes beyond a simple stimulus-response device by implementing cognitive canalizations as follows. Conditions, actions, and schemata can be mutually embedded resulting in a hierarchical arrangement. Since conditions can also be part of a sequence, they act as checkpoints for determining whether the anticipation embodied by the schema is still on the right track. This means that action sequences (at a lower level) are carried out as long as (higher-order) conditions do not veto it. The layers can be stacked upon each other such that schemata refer to the working of other schemata in an increasingly abstract fashion. Consequently, abstract knowledge emerges at the outer boundaries of hierarchical

assemblies but is ultimately embodied as it depends on ontogenetically older lower-level sensorimotor elements.

The features of the algorithm can be summarized as follows. The "semantics" of the cells that the cognitive apparatus works on is not projected from outside into the system. Rather, the system actively produces predictive hypotheses the validity of which is tested against external states. Since schemata do not regard data that lies "by the wayside" (i.e., irrelevant data that is not a priori represented in schemata), the system's overall knowledge acquisition is both accelerated and canalized. In contrast to other programs, this system introduces a novel algorithmic framework for computational discoverers, which neither rely on massive data extraction nor suffer from obstacles such as the frame problem.

Conclusion

I described closed-loop discovery systems as an opportunity to go beyond the cognitive limitations of human scientists, and as a completion to scientific research groups, which are hampered by an administrative and social overhead.

In order to make such scientific artifacts possible, I advocate two major concepts that need to be taken into consideration. (1) *Projective constructivism*, i.e., the reversal of the flow of "information-processing" where the flood of sensory data is not (algorithmically) compressed but projected onto by prior mental structures; (2) *cognitive canalization*, i.e., the claim that the apparent cognitive limitations of the human mind are an expression of its canalizations. Both concepts make artificial closed-loop discovery systems feasible in the sense that they reduce computational requirements and implement embeddedness in the respective domain of reasoning.

References

Appel, K. and Haken, W. 1977, The solution of the four-color-map problem. *Scientific American* 237: 108–121.

Barabási, A.-L., 2002, *Linked*, Perseus Publishing, New York.

Bettoni, M. C., 1997, Constructivist foundations of modeling: A Kantian perspective, *International Journal of Intelligent Systems* 12: 577–595.

Bridgman, P., 1927, *The Logic of Modern Physics*, MacMillan, New York.

Bryant, C. H., Muggleton, S. H., Page, C. D. and Sternberg, M. J. E., 1999, Combining active learning with inductive logic programming to close the loop in machine learning, In: *Proc. AISB'99 Symposium on AI & Scientific Creativity*, pp. 59–64.

Carruthers, P., Stich, S. P. and Siegal, M., eds. *The Cognitive Basis of Science*, Cambridge University Press, Cambridge.

Chase, W. G. and Simon, H. A., 1973, Perception in chess. *Cognitive Psychology* 4:55–81.

Chater, N. and Vitány, P., 2003, Simplicity: A unifying principle in cognitive science?, *Trends in Cognitive Science* 7:19–22.

Dennett, D. C., 1984, Cognitive Wheels: The Frame Problem of AI. In: C. Hookway (ed.) *Minds, Machines, and Evolution: Philosophical Studies.* Cambridge University Press: London.

Giere, R. N., ed., 1993, *Cognitive Models of Science*, Minneapolis: University of Minnesota Press.

Gosselin, F. and Schyns, P. G., 2003, Superstitious perceptions reveal properties of internal representations, *Psychological Science* 14:505–509.

Harnad, S., 1990, The symbol grounding problem. *Physica D* 42:335–346.

Hayes-Roth, F. (1983) Using proofs and refutations to learn from experience. *Machine Learning*, Michalski, R. S., Carbonell, J. G. and Mitchell, T. M., eds., Tioga Publishing: Palo Alto, CA, pp. 221–240.

Hertz, H., 1994, *The Principles of Mechanics*, Dover, New York. German original published in 1894.

Kant, I., 1991, Preface to the second edition, in *Critique of Pure Reason*, translated by Meiklejohn, J. M. D., Continuum, New York, pp. 3–20. German original published in 1787.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B. and Oliver, S. G., 2004, Functional genomic hypothesis generation and experimentation by a robot scientist, *Nature* 427:247–252.

Kozhamthadam, J., 1994, *The Discovery of Kepler's Laws.* University of Notre Dame Press, Notre Dame.

Langley, P., Simon, H., Bradhaw, G. L. and Zytkow, J. M., 1987, *Scientific Discovery: Computational Explorations of the Creative Processes*, MIT Press, Cambridge MA.

Mach, E., 1897, *Contributions to the Analysis of the Sensations*, Open Court, Chicago. German original published in 1886 as *Analyse der Empfindungen*, Gustav Fischer, Jena.

Maturana, H. R. and Varela, F. J., 1980, *Autopoiesis and Cognition*, Reidel, Boston.

Neisser, U., 1976, *Cognition and Reality*, W. H. Freeman, San Francisco.

Piaget, J., 1954, *The Construction of Reality in the Child*. Ballantine: New York. French original published in 1937.

Riegler, A., 1994, Constructivist artificial life, in: Workshop on genetic algorithms within the framework of evolutionary computation, J. Hopf, ed., *Max-Planck-Institute Report* No. MPI-I-94–241, pp. 73–83.

Riegler, A., 1998, "The end of science": Can we overcome cognitive limitations?, *Evolution and Cognition* 4:52–62.

Riegler, A, 2001a, The cognitive ratchet: The ratchet effect as a fundamental principle in evolution and cognition, *Cybernetics and Systems* 32:411–427.

Riegler, A., 2001b, Towards a radical constructivist understanding of science, *Foundations of Science* 6:1–30.

Riegler, A., 2002, When is a cognitive system embodied? *Cognitive Systems Research* 3:339–348.

Searle, J. R., 1980, Minds, brains, and programs. *Behavioral and Brain Sciences* 1:417–424.

Simon, H. A., 1969, *The Sciences of the Artificial*, MIT Press, Cambridge MA.

Skinner, B. F. , 1948, 'Superstition' in the pigeon. *Journal of Experimental Psychology* 38:168–172.

Wigner, E. P., 1960, The unreasonable effectiveness of mathematics in the natural sciences, *Communications on Pure and Applied Mathematics* 13:1–14.

Wolff, J. G., 1982, Language acquisition, data compression and generalization, *Language and Communication* 2: 57–89.