Henriette Haas, Manja Djordjevic and Ann van Ackere

Tarski and the intricacies of reasoning under uncertainty

1. Reasoning under uncertainty

In school we only learn to reason by deduction or under the condition of a finite set of solutions, because tasks are constructed in such a way as to grant a just evaluation of students. In real life many decisions must be taken under uncertainty, as the world is ruled by an infinite number of influences and random events.[1] Work situations present many tasks in which one can fail in spite of the best analytical mind and the sincerest effort. The opposite can also come true: People are successful by sheer luck. Reasoning and decision-making in criminal investigations, namely in the pretrial phase when the ideas about what has happened are still vague, are paradigmatic for all case work, be it in medicine, psychology, management, economics, history, or journalism. The raw material treated in cases contains photos, drawings and descriptions of facts (reports, depositions, letters, etc.) revealed in the context of a questioned incident. What has been detected may be incomplete and may contain random noise, traces that are unrelated to the incident. Analyzing images and texts as evidence is mostly done by simple reading and looking at the pictures, taking notes about whatever salient characteristic the reader happens to perceive. The analysis of

[1] Michael Pidd: Tools for thinking. Modelling in management and science, Chichester ^{3}2009, 31, 46.

the available materials of evidence is an ill-defined task, indeed. Where does one begin? What procedure must one follow? When is the intelligence gathered from the material sufficient for a decision about tracks to be pursued as leads, and which tracks seem to be less promising?

The ultimate *probandum* consists of a larger narrative about what happened in terms of antecedents of the situation as well as human, natural and technical influences on it during a certain timespan. It is a deterministic sequence of causes and influences within a large set of unknown random events. The aim is to explain the detected pieces of evidence of a case in the best possible way. With a series of experiments in different settings we looked at the way case-analyses are being executed in practice. A series of simplified recipes about interpretation are commonly called upon in work situations: (1) distinguishing «facts» from «interpretation»; (2) constructing working hypotheses; (3) determining if the «facts» match the «hypothesis» or not; (4) «improving successively» the working hypotheses within the intelligence cycle. Unfortunately, these common rules of analysis contain many pitfalls.

This study aims at providing illustrations for the intricacies of abduction under complete uncertainty, so as to go beyond the natural way of doing things and offer some insight into how a more proficient analysis might be accomplished. Thus, the above-mentioned naïve routine for case-interpretation needs to be replaced by a better heuristic procedure. Then again practitioners rarely have the time to enter a complex philosophical debate on epistemology. Most of them prefer a set of bullet-proof instructions. This study – including the test case – intends to provide a better set of critical thinking rules for case analyses and serve as materials for educational purposes.

1.1 Questioning common recipes

Unfortunately, there is no pure perception, free of any previous assumptions. The underlying error of this postulation is called naïve realism. The idea of working hypotheses reflects Russell's notion of the hypothesis as a tool to separate plausible from implausible assumptions: «Our theory of truth must be such as to admit of its opposite, falsehood. […] the truth or falsehood of a belief always depends upon something which lies outside the belief itself».[2] Furthermore it is customary in casework to confront all preliminary inferences to the available *observanda*, to see how well they fit. From today's epistemology this procedure is based on a naïve understanding of correspondence theories of «truth».

In a first step naïve correspondence theories can be improved by introducing their most advanced version, Tarski's theorem and its consequences.[3] It postulates a first cognitive level (the so-called «object level») at which facts are observed, and propositions about these facts are formulated. A second level (the meta-plane) consists of a truth-function T about the object level. A proposition p about facts F (represented as *'F'* by some kind of language corresponding to signs) is logically considered «true» if and only if it corresponds to the facts F, so:

(T): $p('F')$ is «true» if and only if F.

Tarski has stripped the semantic notion of «truth» of all its former absoluteness and added precision to the correspondence theories. Propositions are not contained in the information given by the facts themselves; they are educated guesses about the hidden causes or structures behind those facts. No proposition p about F can be called «true» if no fact F can be established. Several propositions p1 – pj about the same set of facts F can all be considered a «match» if they correspond to all (known) facts. «False» propositions contradict some of the facts F or do not explain all of them. This semantic theorem is not uncontested: It does not distinguish between facts

[2] Bertrand Russell: The problems of philosophy, Oxford 1912, chap. 12: Truth and falsehood, 89.

[3] Alfred Tarski: Die semantische Konzeption der Wahrheit und die Grundlagen der Semantik, in: Gunnar Skirbekk (Hg.): Wahrheitstheorien, Frankfurt a.M. 1977, 140–188 (143–145).

and the information contained in them and it only compares one theory with another.[4]

Facts (in the Latin sense of *facta*) consist of documented traces of the past which were collected to analyze a given case. According to the basics of semiotics, the *observandum* (a fact) should be analyzed by referring to units of perception. Those are called signs of evidence. A sign may be a carrier of information. It consists of an outer appearance (its form), which transports possible inner meanings.[5] It is important to recognize that the level of naïve realism cannot be skipped. What can be perceived on a document has to be made explicit. But then such (seemingly) direct signals arriving to eyes and ears need to be enhanced within a next layer of theory which is semiotics. In semiotics Tarski's notion of a «truth function» is replaced by Eco's warning that signs can only be interpreted within their context.[6] In open-ended problems this context is not given by a fixed set of *observanda*, it needs to be reconstructed itself.

The reasoning problem under complete uncertainty can be captured by two null-hypotheses, often taken for granted in practice. With the two working hypotheses we intended to test practitioners work experimentally in order to illustrate what results from naïve recipes concerning the art of interpretation:

H01: Inferences that seem to correspond to the available *observanda* («matches») are helpful in the way of finding more clues to the presumed (yet unknown) real events;

H02: Inferences seeming to contradict the *observanda* («non-matches») will not provide further insights into the presumed real events and can be dismissed.

[4] Another criticism holds that Tarski's theorem fails to address causality. Unfortunately, there is no room here to develop a philosophical debate of correspondence theories.

[5] Charles Sanders Peirce: Pragmatism and pragmaticism (1931), in: Collected papers of Charles Sanders Peirce, V, Cambridge MA 1978.

[6] Umberto Eco: Zeichen. Einführung in einen Begriff, Frankfurt a.M. 1977, chap. 5.19.

As part of the above-mentioned routines, many experts recommend using the intelligence cycle.[7] As an iterative process of finding new materials and explaining what has been observed, conjectures get gradually adapted until they fit the evidence.[8] Let us see what this kind of reasoning typically amounts to. People, having a vague notion of a continuum ranging from totally false to one hundred percent true, tend to think that a proposition is modified step by step with updates so as to contain ever more «truth» and less «falsehood» than before. Each revised proposed theory $pi+1$ supersedes pi and considers another set of facts than its predecessor.[9] Practitioners working the case will judge the plausibility of pi by examining all possible consequences derived from it. As soon as new evidence $Ei+1$ is found, pi can either be confirmed as $pi = pi+1$ which strengthens the belief in it, or it can be overturned or modified in the light of $Ei+1$, to be replaced with an updated new theory $pi+1(\text{«}Ei+1\text{»})$. Along this process older propositions and theories will be rejected and go into a «garbage can» set of bad matches $\neg Mi$. At stage i, they would be considered as «mistaken» in natural language while some pieces of evidence or signs would be considered as random noise or artifacts. Thus, Ei can contain elements that are dismissed in $Ei+1$. Many people believe naïvely that this process would finally converge to «finding out» the unknown «ground truth» GT after a finite number of n iterations, resulting in all the necessary evidence En to terminate the process: $GT = pn(\text{«}En\text{»})$.

Unfortunately, this common heuristic amounts to Popper's idea of verisimilitude,[10] defining a theory's content as a class of logical consequences resulting from it. These should be divided into the truth content (Ct) and falsity content (Cf). The verisimilitude V of a

[7] Olivier Ribaux, Amélie Baylon, Claude Roux, Olivier Delémont, Eric Lock, Christian Zingg, Pierre Margot: Intelligence-led crime scene processing, Part II: Intelligence and crime scene examination, in: Forensic Science International 199/1 (2010) 63–71 (68).

[8] David A. Schum: The evidential foundations of probabilistic reasoning, Evanston IL 2001, 45.

[9] Ibid., 71, 463–468.

[10] Karl Popper: Conjectures and refutations (1963), London 2008, 317.

proposition p_i was calculated as the difference between its truth content and its falsity content. The formula $V(p_i) = Ct(p_i) - Cf(p_i)$ was intended to measure the value of rival propositions. As we tend to associate verisimilitude with the (measurable) type-I error in significance testing of statistical hypotheses, the idea seems plausible, but this a misunderstanding. Verisimilitude combines truth with the content of a proposition whereas probability combines truth with randomness, that is lack of content, of a proposition.[11] In 1974 Miller and Tichý – independently from each other – found that the truth content *Ct* as a set cannot be separated from the falsity content *Cf*. Thus, the falseness of a hypothesis cannot be measured.[12] Every attempt to quantify or to estimate it, may lead to confusion. Incidentally, Popper's work on verisimilitude is in itself an example of an important hypothesis that is formally false while successfully leading to a valid discovery.

If we accepted Miller's and Tichý's result as a basis for practice, no rejection of any proposition would ever be possible in criminal investigation (and other fields). Skepticism is necessary for any scientific endeavor, but it provides no practical instructions regarding by what to replace the criticized approach. In work situations we cannot do without using the iterative verisimilitude approach as a heuristic tool. Categorizing ideas represents common reasoning in teams, although experienced practitioners warn young colleagues to not dismiss working inferences lightly and to always search for alternative explanations. For as long as the inquiry can produce significant amounts of relevant new evidence for any active working hypothesis the intelligence cycle is useful. The very process of strengthening and weakening of different conjectures with the discovery of formerly unknown facts is in itself evidence for as long as the inquiry is correctly conducted into all directions. However,

[11] Ibid., 322.

[12] David Miller: Popper's qualitative theory of verisimilitude, in: The British Journal for the Philosophy of Science 25/2 (1974) 166–177.
Pavel Tichý: On Popper's definitions of verisimilitude. The British Journal for the Philosophy of Science 25/2 (1974) 155–160.

when the investigation stalls, or when it is conducted only into one single direction, then the flaws of these heuristics become manifest.

1.2 Training in the perception and the analysis of signs of evidence

Schum pointed out that «the success […] in generating new and important hypotheses depends to a great extend upon how well we have marshaled or organized the thoughts and the evidence we have».[13] Unfortunately, many analysts fall into the trap of adopting their own first guess or some arbitrary piece of contextual information as a reference point. Cognitive biases are a well-known Achilles' heel and compromise the experts' objectivity. In order to improve perception and interpretation of pictures and written statements some participants learned to apply a heuristic named Systematic Analysis (SA) consisting of five critical thinking rules derived from epistemology, forensic science and cognitive psychology.[14] Comparable to a microscope it can only treat one *observandum* at a time, not great quantities of material.

I. Find schemata and models for the *observandum.*
II. Observe formal aspects of the signs, not only their presumed contents.
III. Dissect the object into its structural components according to models, observe each one.
IV. Note inconsistencies, anomalies, and contradictions.
V. List what seems to be missing or superfluous according to models.

[13] D. A. Schum: The evidential foundations of probabilistic reasoning, 451.

[14] Henriette Haas, Patrick Tönz, Jutta Gubser-Ernst, Maja Pisarzewska Fuerst: Analyzing the psychological and social contents of evidence – experimental comparison between guessing, naturalistic observation and systematic analysis, in: Journal of Forensic Sciences 60/3 (2015) 659–668.

Those rules address all dimensions required by Walker as a minimum for any theory about uncertainty,[15] namely linguistics with rules I., II. and III., causality with rules I., III. and V. and logic with rule IV. The condition to declare one's standpoint added by Schum,[16] is fulfilled by using rule I. to identify models and schemata for the *observandum*. In practice the procedure – when applied to selected potentially meaningful pictures, documents or texts – collects the bricks to construct an inference network later in the investigation. It can point to questions to ask suspects or witnesses; and it can show new promising tracks. Finally, it may offer circumstantial evidence about *mens rea*, about intellectual or professional capacities, psychiatric symptoms and other psychological aspects of a person. It should be mentioned that some of the advice contained in the five rules has been published earlier.[17]

2. Method: comparing analyses with the ground truth

Knowing the ground truth is only possible in an experimental situation when tasks can be constructed around some given well documented event. Only the experiment can provide insights into the pitfalls of real-life situations. In order to find out more about the interpretation of texts and pictures several groups of students and professionals received six test cases to analyze, either before being trained with the five rules-algorithm or after that. The first question of the study to resolve is how to classify their answers. It is the basic question of how to judge working hypotheses: what is useful and what is not?

[15] Vern R. Walker: Theories of uncertainty, in: Marilyn MacCrimmon, Peter Tillers (eds.): The dynamics of judicial proof, Heidelberg 2002, 197–236 (204).

[16] D. A. Schum: The evidential foundations of probabilistic reasoning, 45.

[17] Ibid., 72, 97, 105.

2.1 The test case

For the purpose of this study we selected the most difficult test case (among the six) which exemplifies the challenges of analyzing complex, open-ended problems. Each case has a ground truth *GT* around which it was constructed.

The elephant drawing

The drawing below is meant to represent an elephant. The little figure beside the foreleg of the animal is supposed to be the author of the drawing himself. On demand of the research-team the participant of a scientific study in 1959 had drawn this picture. He was an adult male who was neither physically nor mentally suffering from impairments or illnesses during the test. What was the matter with this man that he drew like this? How would you defend your hypothesis?

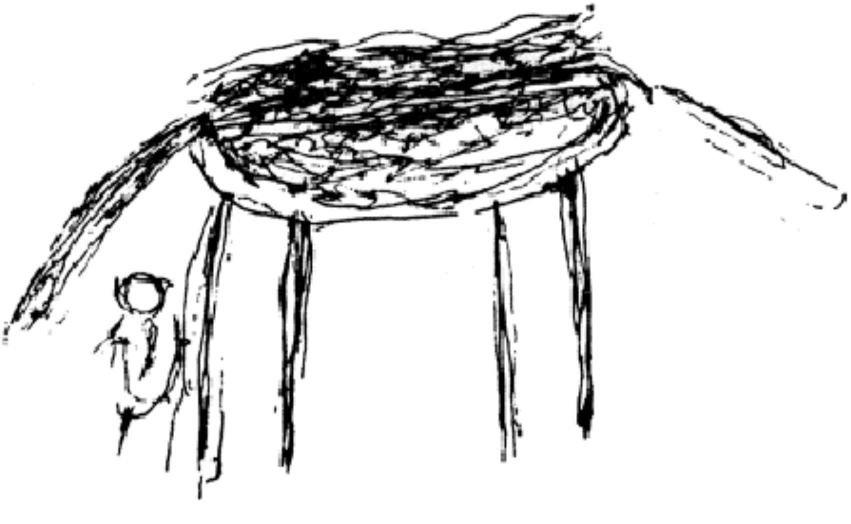

Fig. 15. Drawing of an Elephant.

© Gregory & Wallace.[18] Reprinted with the permission of the authors

[18] The reference is revealed in footnote 25, so as not to spoil the riddle here already.

2.2 Participants

A total of $N = 259$ participants took the elephant task: $n = 155$ were psychology students (aged $M = 26.5$ years [$SD = 5.6$], female : male ratio = 79:21, with $M = 8.6$ semesters [$SD = 3.0$], 48% untrained vs 52% trained), $n = 30$ were law students (aged $M = 24.9$ years [$SD = 3.3$], female : male ratio = 63:37, with $M = 8.1$ semesters [$SD = 2.0$], 40% untrained vs 60% trained) and $n = 74$ were Court and Prosecution professionals (aged $M = 37.1$ years [$SD = 6.6$], with $M = 6.7$ years of work experience in criminal investigation [$SD = 6.7$], female : male ratio = 60:40, 42% untrained vs 58% trained). Two additional psychology students had heard about the drawing and were excluded. The first 80 participants (only psychology students) had 30 minutes to analyze the elephant riddle, whereas all other participants had only 14 minutes (students $n = 105$ and professionals $n = 74$).

2.3 Rating the inferences' success in finding out the ground truth

Participants' working propositions h about a test-case can be evaluated with respect to their matching the facts from the ex-ante perspective but also according to their relevance for detecting the GT from hindsight (ex-post). Those can never be distinguished in real life – according to the reasons for skepticism. Both perspectives can be divided into good and bad answers. Under Tarski's a correspondence or match ($h \in M$) is good while a non-match ($h \in \neg M$) is bad, but according to hindsight a relevant answer ($h \in R$) is useful and an irrelevant one ($h \in \neg R$) is worthless. Non-matches were qualified with the downward modal expressions «implausible» and «impossible». Good matches were qualified with the upward modals «plausible» and «necessary». The latter can only be based on universal laws.

Relevance is defined as the dependency of the ground truth GT under the hypothesis h (h = any hypothetical proposition): h is relevant for GT if and only if $Pr(GT \mid h) \neq Pr(GT)$. Relevant inferences are the *GT* itself (that is the ultimate *probandum*), some partial explanations of the *GT*, and eliminations of what did not happen.[19] Table 1 contains the Cartesian product of ratings.

Table 1

Cartesian Product of Propositions h Between Correspondence to the Evidence M and Relevance R

Relevance *R* Match *M*	Relevant propositions: $h \in R$ $Pr(GT \mid h) \neq Pr(GT)$	Irrelevant propositions: $h \notin R$ $Pr(GT \mid h) = Pr(GT)$
h ∈ M	Good Matches	
modal: necessary	· non-misleading correct eliminations · misleading correct eliminations · necessary partial explanations (*PE*)	*
modal: plausible	· ground truth (*GT*) · broadsides (contains *GT*)	· mistaken but plausible full explanations (FE) · irrelevant partial explanations (seeing ghosts in random noise)
h ∉ M	Bad Matches	
modal: implausible	· implausible but relevant full explanations (*FE*) · implausible eliminations of relevant hypotheses (equivalent of dismissed *GT*/*PE*)	· irrelevant implausible full explanations *FE* (bad observations) · irrefutable ideas
modal: impossible	*	· pseudo-reasoning (speculation)

Legend

Ω Set of all propositions *h*

$M \subset \Omega$ Subset of all matches (necessary, plausible) based on evidence, with $M \cup \neg M = \Omega$

$R \subset \Omega$ Subset of all relevant inferences (necessary, plausible, implausible) considering *GT*, with $R \cup \neg R = \Omega$

[19] D. A. Schum: The evidential foundations of probabilistic reasoning, 172, 96.

The intersection of *M* and *R* contains all relevant propositions *h* which match. $M \cap R$ includes the ground truth *GT*, all correct eliminations plus all partial explanations which show certain aspects of the *GT*. In this category we also counted good propositions that participants had stated and later dismissed, if they were relevant and plausible. All propositions *h* outside of *R* but inside *M* seem plausible ideas ex-ante, yet mistaken in retrospect ($h \in M$ AND $h \notin R$). The classification of specific inferences given by the participants will be shown in Table 2.

Starting with the description of the rating process this study intends to shed light on the difficulties of the evaluation of working inferences. Sentences in natural language are often ambiguous or vague.[20] For example, not all answers were stated as falsifiable sentences. Participants' answers going into the right direction were given half points. The answering sheets provided an observation-section and a section for the final (best) working hypothesis. Often participants would not separate complex inferences from simple descriptions, or they would state several hypothetical scenarios instead of one, or else they did not respect the correct sections. When several inferences were written in the hypothesis section, we classified all of them according to Tarski and relevance. False inferences mentioned only in the observation section were not counted because, following the instructions, we can assume that the participant had rejected them. But we did count additional matching or relevant conjectures in the observation section. This procedure granted that those participants who had decided on one or more favored inference (as required) received less wrongness points than those who had violated instructions by writing only into the observation section without specifying their best guesses.

Another problem of interpretation is that «inferences can usually be decomposed to different levels of granularity, and we are often faced with a choice about the level of detail at which the analysis will be made».[21] At the extreme end of granularity, when

[20] Ibid., 263–265.

[21] Ibid., 5 (see also 90, 492).

the material is dissected into minute pieces (used for any human communication) and recomposed with an inference network of subjective meaning, the error of discovering ghosts within random sequences can occur (as the «Bible-code» fallacy exposed in 1999).[22] In this study we treat only inferences which offer an explanation going beyond a mere description of the visible details. Those will be addressed in another study. According to previous experimentation the application of the five rules does increase the amount of details observed considerably and improve the abductions slightly.[23]

Trying to reproduce the ex-ante situation by classifying inferences according to how well they fit the available case material, showed that a distinction between good and fair matches was fuzzy. Participants' guesses often implied social and natural scenarios consisting of several nested propositions. Some would match the facts seamlessly, others only almost. Some matches were inferences lacking specificity. They were «too bulky on the body of facts», so we called them «broadsides». At the end we chose the best fit among all matches, the one that explained every detail, and would constitute the lead in a real investigation. This was fairly easy. Many participants excluded scenarios seemingly plausible at first sight but not matching the facts after careful observation. These eliminations were deductions from the laws of nature and were counted among matches. Next, we found partial explanations. They explained one part of the *observandum* as a necessary consequence of the laws of nature (for instance «the man has never seen an elephant in his life»). Partial explanations did not attempt to explain some other important aspects of the case (why not draw the head?). In other test cases (not discussed here) subjects would draw inferences based on statistics, not on the evidence at hand. This was qualified as a coincidental relevant match. Finally, some participants explicitly dismissed a good idea in the course of their analysis (dismissed hits

[22] Brendan McKay, Dror Bar-Natan, Maya Bar-Hillel, Gil Kalai: Solving the bible code puzzle, in: Statistical Science 14/2 (1999) 150–173.

[23] H. Haas et al: Analyzing the psychological and social contents of evidence.

or partial explanations). All inferences providing at least some matching aspect were included in the set of good matches *M.* The complementary set of bad matches was easier to define: They were marked by a lack of close observation of the material, contradicted everyday experience and contained many speculative elements. Farfetched speculations were counted as impossible matches when participants overelaborated on their main inference and added wild ideas unrelated to any facts (for instance «the man was afraid of the cold war»). Tautologies and arguments beside the point did not occur at all. Vagueness (for instance «he wanted to draw something else») and all-encompassing inferences are non-refutable ideas according to Russell.[24] We counted them among irrelevant bad matches. All in all, the demarcations between categories are not very clear-cut and would raise endless discussions in practice.

The hindsight (ex-post situation) contains relevant versus irrelevant inferences with respect to the *GT*. The first category captures the ultimate *probandum*, the hits of the ground truth, obviously good matches ($h \in R$ AND $h \in M$). In other test cases some participants had the good idea of the *GT* but dismissed it later. Then there was the category of plausible matches ex-ante which turn out to be untrue ex-post. Thus, promising tracks can be misleading. We did not detect any irrelevant partial explanations in this material, but theoretically they can be made in good faith as explanations for random influences affecting the case's evidence. Finally, we divided the bad matches into relevant and irrelevant ones. If a badly matching idea can be relevant indeed, then the exclusion of this idea is a misleading elimination though it is correct in itself. The double-negation category «badly matching elimination» is an equivalent of a dismissed hit or dismissed correct partial explanation, thus relevant. An implausible elimination that is irrelevant is the same as a farfetched speculation.

[24] B. Russell: The problems of philosophy, 89.

2.4 Inter-Rater-Reliability of the product between matches and relevance

To examine the reliability of the Cartesian rating schema proposed by the first author, she and the second author independently rated a total of 50 participants' analyses of each test case. We measured the reliability of ratings with Crohnbach's alpha (standardized). Statistics were calculated with SAS and are reported in Table 2.

3. Results

Case study results are often presented from hindsight mentioning errors and pitfalls they encountered during the investigation in retrospect. Here we present results so as to enable readers to experience both the challenge of the ex-ante perspective and the illusory easiness of the case-solutions from ex-post. Analysis 1 illustrates the frequent lack of precise descriptions despite 30 minutes time to ponder over the problem. Not astonishingly the hypothesis is vague and poorly adapted to what can be detected in the picture.

Analysis 1 done by an untrained ♀ student (30 min.)

Observations

Very abstract drawing, hardly recognizable for what it is. As if a child had done it. Elephant has very long legs. The man is hardly recognizable. You don't know what is the elephant's fore side and what its hind side. Looks like the man is holding the elephant. The elephant seems very powerful, like a monster.

Hypothesis

Looks as if something oppressing was on his mind, something above him exercising power over him.

3.1 Revealing the ground truth and an optimal analysis

Analysis 2 came closest to the truth. The author immediately dipped into her knowledge of patterns in clinical neuropsychology, however without delivering any description.

Analysis 2 done by an untrained ♀ student (14 min.)

Observations

- *Maybe he has bad imagination and cannot draw very well in general.*
- *Mental status normal: either this man is indeed totally healthy, or else something has been overlooked. When it comes to agnosias they are often hard to discover in clinical exams even today. Agnosia can cause problems with drawing, because patients cannot image and recognize objects (depending on which kind of agnosia).*
  → in 1959 there was not enough research about it (about agnosias).
- *Maybe he was an artist with a special technique.*
- *Or he has never seen an elephant, knowing only that it is big (2nd hypothesis) (and is bad at drawing).*
- *Maybe he has bad imagination and cannot draw very well in general.*

Hypothesis

He may have a neurological condition (for instance agnosia) not known enough to diagnose in 1959.

No one has ever found the key to this puzzle; a rare scenario indeed. In 1959 a middle-aged man who had been congenitally blind and had never seen an elephant in his life underwent surgery for cataracts and recovered full sight. Obviously, he was inexperienced at drawing. Before taking him to the zoo a team of cognitive psychologists visited him for their research on perception and schemata and asked him to make a drawing of the animal.[25]

[25] Richard L. Gregory, Jean G. Wallace: Recovery from early blindness – a case study, in: Experimental Psychology Society Monograph 2 (1963) 65–129 (96).

Congenitally blind people often suffer from neuropsychological dysfunctions and autistic traits,[26] but in 1959 there were no tests available to measure them, therefore he was considered undisturbed then. Here we present a model solution (by the authors):

Model Analysis of the elephant case according to the SA-algorithm

Observations

1) *Comparisons with scientific drawings, children's, cave-men's, patients with different disorders and persons under the influence of drugs.*
2) *Pencil-drawing on white paper with an insecure hand, there is no background and it lacks perspective. Spaces were filled by crude and strong pencil strokes back and forth.*
3) *The elephant consists of a rump, four legs sticking out, no feet, a trunk in front and a tail in its back. Both look quite the same. The animal's head (skull, face, ears, tusks) is missing altogether. The little man consists of a rump and a head, limbs are only partially sketched. The man's face is an empty circle.*
4) *Even though this is a most primitive drawing, the elephant's limbs are connected to its rump, the lines filling out the rump do not trespass the outer limits of the animal and the proportions between the heights of the man and the elephant are more or less correct.*
5) *The missing head is bizarre, it reminds of drawings made by patients with severe neurological problems*

Hypothesis

It seems plausible that the man has never seen an elephant and that he cannot draw (apraxia). Severe deficits in the drawing would suggest that he is blind, but this contradicts the facts. How could behaving like a blind man and yet seeing at least contours be put into a synthesis? Could the evidence stem from mingling effects of different events? What was diagnostically known when the information that he was healthy was put on paper?

[26] R. Peter Hobson, Anthony Lee, Rachel Brown: Autism and congenital blindness, in: Journal of Autism and Developmental Disorders 29/1 (1999) 45–56.

3.2 The panoply of participants' answers

Given the fact that we had to consider more than one single (best) conjecture per person, we calculated the statistics resulting in a set of 609 inferences (Table 2).

Table 2
Working Hypotheses About the Elephant Drawing

	Inferences		*Inter-Rater-Reliability*
Category	*% of all inferences*	*Mean per person*	*Crohnbach's alpha (stand.)*
Relevant inferences ($h \in R$)	67.2%	1.58	0.92
Hits of the ground truth ($h \in M$)	0.4%	0.01	1.00
1) Recovered from early blindness after surgery	0.0%	0.00	–
2) Neuropsychological syndrome, undiagnosed in 1959	0.4%	0.01	1.00
Broadsides ($h \in M$)	0.0%	0.00	–
Falsely dismissed hits of the ground truth ($h \notin M$)	0.0%	0.00	–
Correct partial explanations ($h \in M$)	58.7%	1.38	0.92
1) Has never seen an elephant	14.8%	0.35	0.93
2) Disoriented or no visual imagination	4.4%	0.10	0.87
3) Very bad at drawing	10.8%	0.25	0.86
4) Badly coordinated, shaky	18.1%	0.43	0.78
5) Proportions between human and elephant seem correct	10.7%	0.25	0.99
Falsely dismissed partial explanations (any of 1–5) ($h \notin M$)	2.2%	0.05	–
Correct eliminations ($h \in M$)	1.7%	0.04	0.78
1) Closed eyes or wrong hand or childlike drawing cannot explain the missing head (non-misleading)	1.4%	0.03	0.86
2) He was not blind (non-misleading)	0.3%	0.01	–
Relevant leads in implausible inferences ($h \notin M$) (false negatives)	4.1%	0.10	0.81
1) He was blind	1.6%	0.04	1.00
2) Had a neurological disorder	2.5%	0.06	0.48
Coincidentally true answer based on statistics ($h \in M$)	0.0%	0.00	–

Table 2 (continued)

Category	*% of infer.*	*Mean*	*Crohnbach's*
Irrelevant inferences ($h \notin R$)	32.8%	0.77	0.82
Irrelevant hypotheses among plausible matches (false positives) ($h \in M$)	11.7%	0.28	0.91
1) Was under drugs or alcohol	2.4%	0.06	0.98
2) Was malingering a disorder	0.8%	0.02	–
3) Wanted to sabotage research	6.2%	0.14	0.90
4) Was an artist (art is free, can be bizarre)	0.8%	0.02	0.86
5) Mouth or foot drawing, technical obstacles	1.1%	0.03	1.00
6) Misunderstandings (e.g. meant a horse, giraffe)	0.5%	0.01	0.65
Irrelevant eliminations, irrelevant partial explanations	0.0%	0.00	–
Irrelevant implausible inferences ($h \notin M$)	19.3%	0.45	0.85
1) Raised in isolation, Caspar Hauser syndrome	0.3%	0.01	–
2) Hated or tortured animals	0.8%	0.02	1.00
3) Psychological or sexual problems, complex of inferiority	6.2%	0.15	0.78
4) Raised in wilderness, other culture, or poverty	2.1%	0.05	–
5) Anxious or seeking protection	2.8%	0.07	0.71
6) Drew like a child would draw, or was told to do so	2.1%	0.05	0.82
7) Blindfolded, stress, sloppy, wrong hand	4.8%	0.11	0.91
Impossible (irrelevant) inferences ($h \notin M$)	0.8%	0.02	0.71
Vague ideas (difficult to judge or to refute)	1.0%	0.02	–
All-encompassing ideas (non-refutables)	0.0%	0.00	–
Good matches (Tarski) ($h \in M$)	72.6%	1.71	0.91
Bad matches (Tarski) ($h \notin M$)	27.4%	0.64	0.84

Set of 609 inferences stated by $N = 259$ subjects

Among the matches (n = 442) the percentage of relevant ones was 84% while among the non-matches (n = 167) it was 23%. Consequently, the odds to pursue an irrelevant dead end were about 3 to 1 for a non-match, while only being 1 against 5 for a match. Among the matching full explanations (n = 74), the percentage of hits of the *GT* was 3%. Thus, the great majority of all plausible full explanations (97%) went in the wrong direction.

Table 2 shows that inter-rater reliabilities were sufficient for the classification of the participants' hypotheses within the Cartesian product between matches and relevance. Both our null-hypotheses H01 and H02 must be rejected on the basis of a counter-example. The elephant riddle illustrates the incommensurability of falseness. The hypothesis that the man was blind when he drew the animal not only contradicts the text of the task, it is in total opposition to the facts visible in the drawing. A completely sightless person could never have done it. Thus, an implausible inference can come close to the truth while being far away from the facts. Compared to the jurists, psychologists had a heightened awareness of difficulties with fine motor skills as a neurological symptom. This contradicted the text; however, one should be aware of the fact, that some information previously established was true in the past but can be outdated in the present.

Then again partial explanations and eliminations reflect good reasoning, taking small steps at a time. Plausible full inferences were sorted in the order of their fitting the facts. A good match, an inference to the best explanation, was the intoxication hypothesis. It provides a simple and historically possible cause according to Occam's razor, and yet it is irrelevant for the *GT*. Another good match was «artist»: There are indeed pieces of art that seem unskilled and bizarre (as for instance Joseph Beuys). The psychological phenomena «malingering» and «sabotage» can show strange patterns but must not. We found the explanation of «misunderstandings about which animal to draw between the experimenter and the subject in 1959» (incidence < 2%) not as plausible as the previous conjectures, only a fair match. This is a compromise to limit the number of categories to a manageable number. Eliminations and partial explanations led to relevant results, but they could not detect the *GT* because this is a bold inference going beyond what can be derived from visible elements by laws of nature.

Implausible or impossible conjectures (non-matches) occurred with an alarmingly high incidence: 42.5% of all participants had stated at least one of them in their hypothesis section (distributed almost equally over all three groups). We attribute this result partly to experimental conditions which allowed only 14 minutes of work and made it impossible to search for more contextual information. It also points at the danger of interpretations based on the omission of clearly observable facts. Typical non-matches would ignore that the elephant's legs were joined to its rump (as well as the little man's limbs) and the adequate proportions. They would not consider the insecure hand, neither the repeated lines, nor the care to fill out the rump, which must have taken some time. Other implausible inferences contradicted everyday experience, such as: Nobody in his right mind would omit to draw a mammal's head, even if they do not know its exact shape; it is the most prominent feature in children's drawings. By the same token, everyday knowledge concerns drawings of primitive men, which are far more skilled than the questioned drawing.[27]

Obviously no single example can justify the application of the intelligence cycle which is the verisimilitude approach. But for as long as it is recognized as a heuristic tool and not a formal logical procedure it seems to be helpful in most cases. Scenarios in which almost all unlikely matches are indeed irrelevant must be a very common type of problem in practice, otherwise the fallacy of the verisimilitude approach would be more widely known.

[27] Cf. work of a medieval sculptor on the Basle cathedral who has never seen an elephant https://de.wikipedia.org/wiki/Basler_M%C3%BCnster (visited 16-01-2019)

4. Discussion

4.1 Applying accurate perception, qualified judgement and ethics to complex problems

Reasoning under uncertainty provides plenty of opportunities to err or to be accused of erring even if one is not. One can present a plausible, well-founded hypothesis, which may later be overturned by new evidence in a totally unexpected way. But one can also be accused by hindsight critics of not seeing an obvious contradiction of the working hypothesis with the facts, even if it were relevant for detecting the unknown truth. Only after learning the ground truth does everything fall into place and the solution seems to be obvious and logical. All of a sudden it seems incomprehensible why professionals did not discover it much sooner.

Some might mistake our results as a permission to say that in the art of interpretation «anything goes» henceforth. Such an attitude disrespects scientific ethics and procedure, namely the requirement that analyses be presented in such a way that they provide intersubjective observability and comprehensibility. While abandoning the naive idea of «the ground truth to be found out», the epistemic paradigm still states that interpretation must be done in a qualified way. Schum postulates that the observer must be honest, objective and have an accurate perception.[28] Thus, conjectures which seem plausible only because they are based on sloppiness and arbitrary judgements or even based on deception (fabrication or omission of relevant data) are considered scientific misconduct. Our results speak for much more diligence. They also demonstrate that it is perfectly legitimate to present an individually preferred hypothesis (even if it is a minority opinion) as the final conclusion about a case, *if – and only if* – those facts, which contradict the conclusion, as well as the alternative explanations are not kept in the dark. Creativity and subsequent investigations must be facilitated and not impeded.

[28] D. A. Schum: The evidential foundations of probabilistic reasoning, 229.

4.2 Structuring the presentation of the intelligence drawn from the *observandum*

The output of a natural observation in our sample was either an unstructured assembly of incomplete sentences and arrows, or else a well-edited text. The latter form of presenting an analysis is nice to read but hard to re-work. Any major modification requires a total re-write (cf. Analyses 1 and 2). A comparison to the work of a trained professional (Analysis 3) shows how much easier it is to criticize and improve the results following the structured algorithm:

Analysis 3 done by a trained ♀ professional (14 min.)

Observations

1) *Drawing made by a «normal» man. Even kids draw an elephant's big ears and trunk, but not this man.*
2) *Black & white drawing, no perspective. Insecure hand (poorly defined contours (→ does not know what he is drawing, is unsure).*
3) *Big animal (?) without head, long limbs. Next to it a person without limbs (no arms and legs). Legs of the elephant too thin, but very long. No head but two tails.*
4) *Contradicts the anatomy of elephants.*
5) *No head, no ears, the person's tie seems superfluous.*

Hypothesis:

The man has never seen an elephant. He knows that it is a big animal. Not more. Maybe he confounds it with a giraffe.

Using the five rules provides a structured procedure and a table of evidence that can be criticized from all angles. Other experts can easily complete or improve a mediocre analysis by adding or criticizing models and schemata used (rule I.) and then find more signs of evidence according to rules II. to V. Reasoning under uncertainty is a matter of discipline; like every successful endeavor in life it is based on one percent inspiration and 99 percent transpiration.

4.3 How to improve reasoning under uncertainty

Statements like «reasoning under uncertainty is an art» are not helpful in practice. What can be derived from the present study to make into a craft? As proposed by Wagenaar, Koppen and Crombag, implicit assumptions about life played an important role in the participants' reasoning process.[29] Our results underscore the necessity to make them explicit by specifying «everyday rules» and schemata used to back up an inference. James Reason mentions the retrieval of incomplete semantic knowledge as a major source of error.[30] Few participants came up with the idea that children never draw animals without their heads. Teachers would certainly testify that they have never seen a healthy child do this, but statistics are unavailable. The decisive role of gathering more contextual knowledge about drawings and about conditions that can influence them is obvious.[31]

We also saw that deductive partial explanations and eliminations seem to be a good way of trying to find relevant answers to unresolved cases. When an investigation stalls, one should not prematurely be fixed on the full explanation of the story of what presumably happened but make small and coherent advances in the interpretation of the details of the evidence with specific sub-hypotheses. Then again one cannot absolutely trust such deductions, as one does not know how the traces arrived there. Were they caused by the presumed incident or by random events?

In order to collect ideas about the so-called big picture, a brain storming about all imaginable scenarios dipping deep into the sources of life-experience should be encouraged.[32] The background

[29] Willem A. Wagenaar, Peter J. van Koppen, Hans F. M. Crombag: Anchored narratives. Psychology of proof in criminal law, London 1993, 61, 232, 235.

[30] James Reason: Human error, Cambridge 1990, 112.

[31] M. Pidd: Tools for thinking, 59, 97; W. A. Wagenaar et al.: Anchored narratives, 237–240.

[32] M. Pidd: op. cit., 60.

knowledge cannot always be found in books.[33] Professionals should introduce their field experience and enrich the arsenal information but, in addition, it is also a wise choice to consult people who know the given social, cultural, economic, technical, natural or work-related context. Ideas collected in the brainstorming process need to be bold – thus go far beyond what it visible in the evidence – otherwise nothing new can be discovered.

4.4 Integrating naïve recipes into hierarchical levels of epistemology

The final task in order to avoid professional errors in reasoning under uncertainty is to provide a deeper understanding of different levels of methodology. The general idea of the critical thinking schema shown in Table 3 is to integrate the historical ideas of «truth finding» or positivism into several layers of reasoning. The schema starts on its lower levels with the everyday recipes and integrates them step by step into a more sophisticated methodology.[34]

Table 3
Levels of reasoning and perception under uncertainty from the bottom-up perspective

4. Skepticism (meta-theory)
Errors can result from a naïve belief in realism, correspondence theories and the intelligence cycle. · An interpretation of signs within their context can be a match and yet be mistaken. · An interpretation of signs can contradict the picture of signs yet come very close to the ground truth.

[33] W. A. Wagenaar et al.: Anchored narratives, 47.

[34] Here we consider only the bottom-up perspective. Its relationship with top-down theories, developed on a set of premises, cannot be discussed here.

3. Separation between plausible and implausible assumptions (theory)	
Positivism (Tarski / Popper): Decision is made on higher level by using a truth function about correspondence.	Semiotics (Eco / Peirce): The meanings of a signs can only be inferred within their context.
2. Critical realism (observation)	
Early positivism (Russell): Hypotheses need to be falsifiable statements tested by something outside them.	Semiotics (Eco / Peirce): · Observation relies on signs as units of perception. · A sign is more than a signal. It consists of an outer form conveying inner meaning(s).
1. Naïve realism (information)	
Direct perception provides information	

It is important to realize that none of these levels can be dismissed. Naïve realism states that what is there on paper under everybody's eyes cannot be ignored. This is called evidence. If obvious signals were omitted, important information about the *observanda* could be entirely lost or certain aspects could be systematically selected with a bias favoring the observer's own opinion (so-called cherry-picking). The result of disregarding facts could also amount to sophistries and *petitio principii*. If the second level of distinguishing signs from signals were not respected, then some up-front meanings could falsely be taken for granted and the formal aspects of signs cannot sufficiently be perceived. If hypotheses were not stated as falsifiable sentences, then vagueness and ambiguity easily set in (as the experiment has shown). They provide no further insights. Allowing vagueness and ambiguity has the side effect to open up the debate to rhetorical pirouettes favoring biases and distorting the picture in the readers' reception. The necessity of a test level for the plausibility of any theory about an incident is also obvious. It distinguishes qualified from unqualified beliefs and provides the grounds for critical discussions and brainstorming. Eco's warning prevents from committing the Bible code fallacy of recomposing signs out of their context in order to create an arbitrary mosaic. Then again, the testing theories-level cannot be the last word about the result of an analysis done under complete uncertainty as has shown our experiment with the elephant case. While the diligence of a given analysis can be established and praised, its results must still be subjected to some skepticism.

Acknowledgements

We want to express our thanks to all those who have participated in our study: Especially to all professionals from MAS Forensics at University of Lucerne, from the Prosecutors office and Courts of the Canton of Zürich, Switzerland, and also to Christiane Trapp, JD for letting her students participate in the experiments. To Marianne Pithon, ès lettres we owe for her close reading of the English.

conexus 2 (2019) 25–51

https://doi.org/10.24445/conexus.2019.02.004

Prof. Dr. Henriette Haas, Universität Zürich, Psychologisches Institut,
Binzmühlestrasse 14 / Box 1, 8050 Zürich
henriette.haas@psychologie.uzh.ch

Manja Djordjevic, M. Sc. Psychologin, Elsässerstrasse 31, 4056 Basel
m-djordjevic@hotmail.com

Prof. Dr. Ann van Ackere, Université de Lausanne, École des hautes études commerciales, Département des opérations, Quartier UNIL-Chamberonne, Bâtiment Anthropole, 1015 Lausanne
ann.vanackere@unil.ch