# Catching cognitive biases in an erroneous decision making process

Valentin Fouillard, Nicolas Sabouret, Safouan Taha, Frédéric Boulanger

# Catching cognitive biases in an erroneous decision making process

Valentin Fouillard[1,2], Nicolas Sabouret[1], Safouan Taha[2] and Frédéric Boulanger[2]

***Abstract*— This paper proposes a logic-based model to study erroneous decision making by humans in accident reports. Our model is based on minimal belief revisions and forward chaining. It computes possible mental states that could explain the operator's behavior. From this sequence, we extract logical patterns that correspond to possible cognitive biases responsible for the erroneous decision making. We apply this model on the crash of Air France's 447 flight in 2009.**

## I. Introduction

Three years after the crash of the Air France flight 447 Rio-Paris, the report of the Bureau of Investigation (BEA in French, for *Bureau d'Enquêtes et d'Analyses*) is published [1]. This report shows that the pilots did not identify the stall situation even though the stall alarm rang over 75 times. From an external point of view, the behavior of the pilots can seem totally irrational. Nevertheless, the BEA outlines a possible confusion between the stall situation and an overspeed situation. Several elements can support this hypothesis: the lack of visual information, wrong indications from the flight directors, irregular stall warning, etc. If we assume that the pilots believe to be in an overspeed situation, all their actions are rational.

In this paper, we propose to use computer simulation to study and explain such situations, where a rational and skilled human operator adopts a behavior that seems faulty and irrational when considering all the information available. Our goal is to determine the possible causes of errors by reconstructing the mental state of the operator, based on his actions and observations (section II). To this goal, we define a logic-based formal model (section III). We rely on belief revision mechanisms to generate the possible mental states. We propose a mechanism to detect *cognitive biases* that can explain the erroneous decision making (section IV). Section V illustrates our method on the AF-447 situation. We show that several explanations are possible for the pilot's behavior, beyond those proposed by the BEA analysis. Using cognitive biases, our model give hints on the plausibility of each scenario. The last two sections of this paper discuss some related work (section VI) and perspectives of our model (section VII).

## II. Overview of the approach

Our goal is to model situations in which an operator adopts an irrational behavior. By irrational behavior, we mean an action (or a set of actions) that are in contradiction with the state of the world. For example, in the crash of the AF-447 flight, the pilots perform a sequence of actions that keeps the aircraft in a stall situation instead of getting out of this situation. To model such situations, we consider four elements, as illustrated in Figure 1:

- a sequence of actions performed by the operator;
- the observations that the operator can make at each step;
- the initial beliefs $\mathrm{B}_0$ of the operator;
- the reasoning rules of the operator.

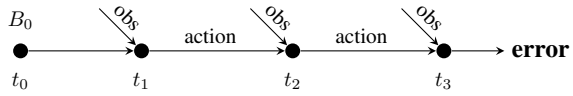



Fig. 1. Initial beliefs, observations, and actions

We define a *rational behavior* as a sequence of belief states that are *logically consistent* with the actions and observations performed at each time step. Two situations can lead to inconsistencies:

- Some new observations are not consistent with the beliefs or deductions of the operator: this corresponds to new information to take into account;
- Beliefs or deductions of the operator are not consistent with the action performed: this corresponds to the irrational behavior situation we want to capture.

Our goal is to restore the consistency all along the succession of actions and beliefs. This is a well-known problem in logic-based modeling, called *belief revision* [2]. It consists in removing some of the beliefs or observations to obtain a consistent subset of propositions. We claim that each revision correspond to a possible mental state of the operator that could explain his erroneous decision. In the example of the AF-447 flight, ignoring the alarm can lead the pilots to believe they are in an overspeed situation, which explains their actions. Our goal is then to decide which belief revisions are acceptable from a psychological point of view.

## III. Model

### A. Syntax

Our model is based on first-order logic with each predicate indexed by a time step. For example:

$\mathrm{clouds}(l)_t \quad \rightarrow \quad$ there are clouds in location $l$ at time $t$
$\mathrm{alarm}_t \quad \rightarrow \quad$ the alarm rings at time $t$

For easier reading, any free variable or temporal index is considered as universally quantified:

$$P(x,y)_t \quad \Leftrightarrow \quad \forall x, \forall y, \forall t \quad P(x,y)_t$$

[1] Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France.
[2] Université Paris-Saclay, CNRS, ENS Paris-Saclay, CentraleSupélec, Laboratoire Méthodes Formelles, 91190, Gif-sur-Yvette, France.
firstname.lastname@universite-paris-saclay.fr

We define the language $\mathcal{L}_0$ with the following grammar:

$$\alpha ::= p \,|\, \neg\alpha \,|\, \alpha_1 \wedge \alpha_2 \,|\, \alpha_1 \vee \alpha_2 \,|\, \alpha_1 \Rightarrow \alpha_2$$

where $p \in \mathcal{P}red$, a set of of temporally indexed predicates. Human beliefs and observations, as well as inference rules for the system, are represented by formulas in the $\mathcal{L}_0$ language. Inference rules are numbered and hold at all time steps. For example, to model the fact that human beings assume a blue sky when there is sun and no clouds at some location $x$, we write:

$$R_1(x)_t \equiv ((\neg\,\mathrm{clouds(x)}_t \wedge \mathrm{sun(x)}_t) \Rightarrow \mathrm{bluesky(x)}_t)$$

Belief revision in our model consists in "ignoring" a belief, an observation or a rule. In the above example, ignoring $R_1$ means that we can't infer a blue sky from sun and no cloud.

We also define a set $\mathcal{A}$ of temporally indexed actions:

$$a_t \quad \rightarrow \quad \text{action } a \text{ is performed at time } t$$

We define the language $\mathcal{L}$ as an extension of $\mathcal{L}_0$ by adding the actions $\mathcal{A}$ to the set of predicates, as well as the two following operators:

$$\phi ::= \alpha \,|\, [\alpha]act \,|\, act :: \alpha$$

with $\alpha \in \mathcal{L}_0$ and $act \in \mathcal{A}$. $[\alpha]act$ states that $\alpha$ is the precondition of action $act$ and $act :: \alpha$ states that $\alpha$ is the effect of $act$. For example:

| | |
|---|---|
| $[\neg\,\mathrm{locked}_t]\,\mathrm{doOpen}_t$ | action doOpen requires that the door is not locked. |
| $\mathrm{doOpen}_t :: \mathrm{open}_{t+1}$ | action open has for effect that the door is opened at the next time step. |

The language $\mathcal{L}$ allows us to write sets of logical propositions that represent irrational behaviors.

### B. Situation description

The description of an irrational situation to be analyzed is composed of several elements:

**The initial beliefs** $\mathcal{B}_{init}$: a set of predicates $p \in \mathcal{P}red$ representing the initial beliefs of the operator.

**The reasoning rules** $\mathcal{R}$: a set of $\mathcal{L}$-formulas with a free temporal index $t$ (they hold at all times). Some of these rules support the deduction of new propositions from beliefs and observations (*e.g.* $(\neg\,\mathrm{clouds(x)}_t \wedge \mathrm{sun(x)}_t) \Rightarrow \mathrm{bluesky(x)}_t)$). Some others define the preconditions and effects of actions (*e.g.* $[\neg\,\mathrm{locked}_t]\,\mathrm{doOpen}_t$).

**The desires** $\mathcal{D}$: a set of positive or negative literals of $\mathcal{P}red$. They represent the operator's goals, *i.e.* things that should be satisfied at the next time step.

**The observations** $\mathcal{O}bs = \{Obs_1, \ldots, Obs_t\}$: each set of observations $Obs_i$ is a self-consistent set of positive ($o \in \mathcal{P}red$) or negative ($\neg o \in \mathcal{P}red$) literals that correspond to the observations at time step $t$. Literals that do not appear in $Obs_i$ are unknown (not observed) at this time step.

**The track** $\mathcal{T} = \{a_1, \ldots, a_t\}$: a sequence of actions of $\mathcal{A}$ representing the actions that the operator has performed at each time steps.

**The permanent rules** $\mathcal{C}$: a set of propositional formulas that hold at all time steps and cannot be ignored in the belief revision process. These rules describe physical properties that cannot be violated.

Our goal is to use belief revision to compute "rational" mental states of the operator, *i.e.* sets of consistent propositions, from the description of the situation..

### C. Mental state definition

A mental state is a set of predicates and rules in the language $\mathcal{L}$ that describes the beliefs, desires, observations and reasoning rules which have been retained by the operator. We define $B_t$ as the mental state at time $t$. The initial mental state of the operator is the conjunction of his initial beliefs, reasoning rules, desires (for the next time step), and permanent rules:

$$B_0 = \mathcal{B}_{init} \wedge \mathcal{R} \wedge \mathcal{D} \wedge \mathcal{C}$$

At each step, $B_t$ contains a subset of $\mathcal{L}$. Everything that is not in $B_t$ is unknown of the operator. Therefore:

| | | |
|---|---|---|
| $\mathrm{p}_{t'} \in B_t$ | $\rightarrow$ | I believe at $t$ that $p$ is **true** at $t'$ |
| $\neg\,\mathrm{p}_{t'} \in B_t$ | $\rightarrow$ | I believe at $t$ that $p$ is **false** at $t'$ |
| $\mathrm{p}_{t'} \notin B_t$ $\wedge \neg\,\mathrm{p}_{t'} \notin B_t$ | $\rightarrow$ | I **don't know** at $t$ whether $p$ is true or false at $t'$ |

In order for the operator to behave in a rational way, our model requires that each $B_t$ is consistent: the set of beliefs of the operator is consistent with his actions, his observations, his desires and his reasoning rules. We also need to determine how the model evolves from a mental state $B_{t-1}$ (consistent) to a new mental state $B_t$ (consistent), taking into account the observations $Obs_t$ and the action $a_t$, which can create inconsistencies. This is the object of the belief revision.

Our first hypothesis is that by default, *the operator continues to believe what he believed at the previous time*: each $B_t$ includes a priori the previous mental state $B_{t-1}$.

Then, we add the observations to this belief base. If these observations do not contradict the beliefs, namely if the set $B_{t-1} \cup Obs_t$ is consistent, there is no problem. Otherwise, a belief revision is necessary to restore consistency.

In a second step, since we know that the operator has performed the action $a_t$ at time step $t$, we need to make sure that his beliefs allow him to perform this action, *i.e* they are consistent with the preconditions and effects of the action. If it is not the case, we need to establish a diagnosis on the beliefs of the operator that can explain the choice of this action, as explained by Reiter [3]. Wassermann [4] shows that this diagnosis problem can also be resolved with the help of a "minimal" belief revision. Therefore, the mechanism of belief revision can be used both to take into account the new observations and to diagnose the actions that seem to be erroneous from the point of view of an omniscient observer.

### D. Belief revision

Belief revision restores the consistency of the beliefs of an agent facing new contradictory information [2]. In our model, this mechanism is the main tool to compute the transition from a mental state to a new one and so to find a possible rational reasoning.

However, not all revisions are equivalent. Consider the following example:

$$\begin{array}{lcl} R_1 & \equiv & (\mathrm{rain}_t \wedge \mathrm{cold}_t) \Rightarrow \mathrm{snow}_t \\ B_0 & \equiv & \{\mathrm{cold}_1, \mathrm{rain}_1, R_1\} \\ Obs_1 & \equiv & \{\neg\, \mathrm{snow}_1\} \end{array}$$

The operator does not observe snow although he believes that it is raining and it is cold (and therefore that it is snowing by application of $R_1$, which is in $B_0$); there is an inconsistency. Any strict subset of $B_0 \cup Obs_1$ is a possible belief revision. Therefore, we can choose to ignore the observation, to give up on one of the beliefs on rain or cold, or to withdraw deduction rule $R_1$. We can also choose to ignore at the same time $R_1$ *and* a belief: from a formal logic viewpoint, there is no reason to choose one over the other. Yet, from a human cognition viewpoint, ignoring all the beliefs for taking into account the new information seem to be excessive.

This problem was studied by Carlos Alchourron, Peter Gärdenfors and David Makinson [5] and resulted in the AGM theory (after the name of the authors). Their proposition is to define a set of axioms that characterize a *minimal* belief revision. In our case, the solution that consists in ignoring both $\mathrm{rain}_1$ and $\mathrm{cold}_1$ is not minimal because eliminating $\mathrm{rain}_1$ or $\mathrm{cold}_1$ only is enough. In other words, we are looking for a minimal correction to bring the system $B_0 \cup Obs_1$ back to consistency. This is known as determining a *Minimal Correction Set* (MCS).

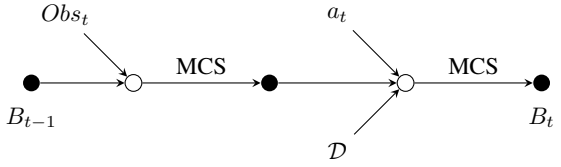
More formally, for a given system $\Phi = \{\phi_1, \phi_2 \dots \phi_n\}$, $M \subseteq \Phi$ is a MCS of $\Phi$ if and only if:

- $\Phi \setminus M$ is consistent
- $\forall \phi_i \in M, (\Phi \setminus M) \cup \{\phi_i\}$ is inconsistent

In our example $\{\mathrm{rain}_1\}$ is an MCS, $\{\mathrm{cold}_1\}$ is another one, but $\{\mathrm{rain}_1, \mathrm{cold}_1\}$ is not one. We use Liffiton's algorithm [6] to calculate the set of MCS of a system. This algorithm has the benefit to take into account a subset of $\Phi$ of beliefs that cannot be ignored. We use this to prevent the belief revision to ignore the action performed by the operator and the permanent rules $\mathcal{C}$.

#### *Application to our model*

The following figure illustrates the computation of the successive belief states of the operator:




First of all, we add the observations. If the result (white dot) is inconsistent, we look for an MCS to make it consistent (black dot). Then, we add the actions and the desires for the next step and we compute a new MCS to perform a diagnosis. As a consequence, the mental state $B_t$ contains all observations consistent with the previous belief state and the action performed at time $t$.

At each step, we don't necessarily obtain a unique MCS. Like in the weather example above, there is more than one possible correction (ignore the rain, ignore the observation, ignore $R_1 \dots$). We therefore obtain a tree structure of states $B_i$, where each branch corresponds to a possible "choice of revision" for the operator. Each path from the root to a leaf defines a possible cognitive behavior. We call this a *scenario*. Not all scenario are evenly relevant: in the next section, we show how to use *cognitive biases* to extract the most plausible ones, *i.e.* those who correspond to a possible behavior of a human operator.

#### *Implementation*

We have implemented our model under the SMT-LIB language using the z3 solver. This allows us to take advantage of the incremental operators "push" and "pop" to query the solver without reloading the model each time. We load the definitions of predicates and rules once, and then push the assertions corresponding to each query as part of our MCS tree construction algorithm.

Moreover we have defined a modeling language with an ANTLR (*ANother Tool for Language Recognition*) grammar, and we wrote a program to translate the models from this language to SMT-LIB. This allow us to write and analyze models in an efficient and systematic way.

## IV. Cognitive Biases

### A. Cognitive biases in social science

Finding a rational reasoning explanation to an irrational behavior has been widely studied in social sciences. Indeed, according to Tversky and Kahneman [7], humans use mind shortcuts (heuristics) to compensate their *bounded rationality*. These heuristics can be very effective, but they can lead to errors. They are the *cognitive biases*. Cognitive biases can lead to predictable irrational behaviors. Our proposition is that these biases offer a plausible explanation to accidents in decision making processes.

Dimara [8] enumerates 151 cognitive biases in the scientific literature. Our work however focuses on specific situations: incidents, crashes, collisions or disasters. In this context, Murata et al. [9] propose a subset of cognitive biases that play a key role in such accidents.

In the following, we give a non-exhaustive set of logical patterns (using our model based on Minimal Correction Sets) that capture biases from Murata's list. Each bias is illustrated with a practical example.

### B. Attention bias

The attention bias is defined as a selection among perceived information, based on concern or emotion [10]. When the operator retains an observation that is inconsistent with one of his desires, then his attention is focused on this

observation. As a result, an MCS representing an attention bias is defined by the following conditions:

1) $p, q \in Obs_t$ — $p$ and $q$ are two observations
2) $\exists d \in \mathcal{D}.\, (B_{t-1} \wedge p) \Rightarrow \neg d$ — $p$ contradicts a desire
3) $q \in \mathrm{MCS}_t$ — $q$ is ignored

Let us consider the following example:

$$\begin{array}{lll} Obs_1 & \equiv & \{\mathrm{alarm}_1, \mathrm{reserve}_1\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{l} R_1 \equiv \mathrm{alarm}_t \wedge \neg\, \mathrm{reserve}_t \Rightarrow \mathrm{outFuel}_t \\ R_2 \equiv \mathrm{alarme}_t \wedge \mathrm{reserve}_t \Rightarrow \neg\, \mathrm{outFuel}_t \\ R_3 \equiv [\mathrm{outFuel}_t]\, \mathrm{land}_t \end{array} \right\} \\ \mathcal{D} & \equiv & \{\neg outFuel\} \\ a_1 & \equiv & \mathrm{land}_1 \\ \mathrm{MCS}_1 & \equiv & \{Obs_1(\mathrm{reserve}_1)\} \end{array}$$

In this example, the operator has the desire not to run out of fuel. He knows that he is out of fuel when the alarm rings and there is no reserve. In spite of not running out of fuel, he decides to make an emergency landing. The MCS shows that the reserve was ignored. The attention bias can be applied here to explain this decision (the pilot focused on the alarm).

### C. Commitment bias

The escalation of commitment is the tendency to persist in an irrational behavior in spite of increasingly negative outcomes [11]. In our model, we consider that a commitment bias is present when an operator rejects the effect of a previous action and keeps performing this action:

1) $\exists d \in \mathcal{D}, B_{t-1} \Rightarrow \neg\, \mathrm{d}_{t-1}, B_t \Rightarrow \neg\, \mathrm{d}_t$
   There exists a desire $\mathrm{d}$ not satisfied at $t-1$ and $t$.
2) $\mathrm{a}_t = \mathrm{a}_{t-1}$ — same selected action $\mathrm{a}$ at $t-1$ and $t$.
3) $B_{t-1} \Rightarrow \mathrm{d}_t$ — Action $\mathrm{a}_{t-1}$ should have satisfied $\mathrm{d}$ (remember that $\mathrm{a}_{t-1} \in B_{t-1}$)
4) $B_t \Rightarrow \mathrm{d}_{t+1}$ — Action $\mathrm{a}_t$ is believed to satisfy $\mathrm{d}$
5) $\mathrm{R}^a_k \in \mathrm{MCS}_t$ with $\mathrm{R}^a_k$ the rule that defines the effect of action $\mathrm{a}$. The operator had to ignore these effects to be consistent with the observations at $t$ (which tell her that $\mathrm{a}_{t-1}$ did not work).

Let us consider the following example:

$$\begin{array}{lll} Obs_1 & \equiv & \mathrm{speed}_1 \quad Obs_2 \equiv \mathrm{speed}_2 \\ \mathcal{R} & \equiv & \{\ R_1 \equiv \neg\, \mathrm{ice}_t \Rightarrow (\mathrm{brake}_t :: \neg\, \mathrm{speed}_{t+1})\ \} \\ \mathrm{a}_1 & \equiv & \mathrm{brake}_1 \quad \mathrm{a}_2 \equiv \mathrm{brake}_2 \\ \mathrm{MCS}_2 & \equiv & \{R_1\} \end{array}$$

In this example a driver tries to brake to reduce its speed, which does not work (probably because he is slipping on ice). Although the decision was not good, he tries again to brake at step 2. The MCS ignores the effect of the braking action, allowing the operator to perform the action despite the inconsistency between the expected effect and the new information.

### D. Overconfidence

Overconfidence is the tendency to consider our own judgment as more accurate and efficient than it is really [12]. In our model, we identify overconfidence when the operator predicts a state of the world and then rejects any information that does not support his belief.

1) $p \in Obs_t$ — $p$ is an observation
2) $(B_{t-1} \wedge p) \Rightarrow \bot$ — inconsistent with the prediction
3) $p \in \mathrm{MCS}_t$ — $p$ is ignored

For example:

$$\begin{array}{lll} Obs_1 & \equiv & \mathrm{clouds}_1 \quad Obs_2 \equiv \neg\, \mathrm{rain}_2 \\ \mathcal{R} & \equiv & \{R_1 \equiv \mathrm{clouds}_t \Rightarrow \mathrm{rain}_{t+1}\} \\ \mathrm{MCS}_2 & \equiv & \{Obs_2(\neg\, \mathrm{rain}_2)\} \end{array}$$

The operator was overconfident in his prediction of rain.

### E. Confirmation bias

The confirmation bias is the tendency to prefer information that confirms our beliefs over information that challenges them [13]. We identify this bias in our model when the operator has the choice between contradictory pieces of information in $Obs_t$ and keeps the elements that confirm his beliefs (*i.e.* the selected observations can be used to derive facts that belong $B_{t-1}$).

1) $p, q \in Obs_t$ — $p$ et $q$ are two observations
2) $(B_{t-1} \wedge p) \not\Rightarrow \bot$ — $p$ is consistent
3) $(B_{t-1} \wedge p \wedge q) \Rightarrow \bot$ — $p$ and $q$ are inconsistent
4) $\exists R_k \in B_{t-1}, ((B_{t-1} \setminus R_k) \wedge p \wedge q) \not\Rightarrow \bot$
   $p$ et $q$ are consistent when $R_k$ is ignored
5) $\exists\ B \subseteq B_{t-1} \left\{ \begin{array}{lll} (B_{t-1} \setminus B) & \not\Rightarrow & B \\ (B_{t-1} \setminus B) \wedge p & \Rightarrow & B \end{array} \right.$
   $p$ confirms some beliefs
6) $(p \notin \mathrm{MCS}_t \wedge q \in \mathrm{MCS}_t) \vee R_k \in \mathrm{MCS}_t$
   either $p$ is prefered to $q$, or $R_k$ is ignored

For example:

$$\begin{array}{lll} Obs_1 & \equiv & \{\mathrm{greenLight}_1\} \\ Obs_2 & \equiv & \{\neg\, \mathrm{redLight}_2, \mathrm{alarm}_2\} \\ \mathcal{R} & \equiv & \left\{ \begin{array}{lll} R_1 & \equiv & \mathrm{greenLight}_t \Rightarrow \neg\, \mathrm{failure}_t \\ R_2 & \equiv & \neg\, \mathrm{redLight}_t \Rightarrow \neg\, \mathrm{failure}_t \\ R_3 & \equiv & \mathrm{alarm}_t \Rightarrow \mathrm{failure}_t \end{array} \right\} \\ \mathrm{MCS}_2 & \equiv & \{R_3\} \end{array}$$

Here the observation of the green light leads to the conclusion that there is no failure at step 1. The observation that there is no red light at step 2 confirms that there is no failure, contrary to the observation of the alarm. The MCS gives more weight to the information that confirms that there is no failure. Note that the MCS $\{Obs_2(\mathrm{alarm}_2)\}$ would also be a case of confirmation bias.

### F. Conclusion

The above examples show that it is possible to define patterns that identify biases in rational reasoning scenarios within the system of beliefs, observations, reasoning rules and MCS. Our proposition is that among all possible scenarios, a scenario that follows cognitive biases is a more plausible explanation for an irrational behavior.

## V. Application to the case of the Rio-Paris flight

### A. Explanation of the situation

To illustrate our methodology, we model a simple version of the Rio-Paris flight crash based on the report of the BEA [1]. Four main devices played a role in this accident:

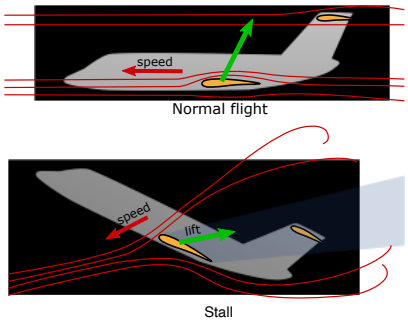



Fig. 2. Stall situation (*source : Caliver, Wikimedia Commons*)

- The Flight Director (FD) that tells the pilot what maneuver to make to reach the programmed trajectory;
- The stall alarm that rings when the aircraft is in a stall situation (loss of lift resulting in a fall at high incidence, see Figure 2);
- The altimeter that gives a *vertical speed* ($V_z$), indicating the fall of the aircraft;
- The Pitot probe, defective at the time of the crash, which indicates the *speed* of the plane (and thus a possible overspeed).

Moreover, the pilots felt a *buffeting* (vibrations of the aircraft) at the start of the incident, which they misinterpreted.

To understand the error of the pilots, one must know that when an airplane is in a stall situation, the correct action is to push the control stick (*i.e* use the flight control to tilt the aircraft forward). On the other hand, when the aircraft is in overspeed, it is necessary to pull the control stick to tilt the plane backward [14]. It was the confusion between these two situations, in the absence of clear instrument indications and without external visibility, that led the aircraft to crash.

### B. Our model of the problem

We can summarize the situation as follows:

**(t=1)** The flight indicators display an abrupt acceleration, the stall alarm rings and buffeting is felt. The pilot pulls the stick.

**(t=2)** The stall alarm and the buffeting stop, the vertical speed increase (loss of altitude) and the flight director requests to pull the stick. The pilot pulls the stick.

**(t=3)** The stall alarm is still off, the vertical speed continues to increase and the flight director is still requesting to pull the stick. The pilot pushes the stick.

**(t=4)** The stall alarm turns back on, the vertical speed increases and the flight director is still requesting to pull the stick. The pilot pulls the stick.

Our model proposes explanations for the behavior of the pilot with the help of cognitive biases and belief revision. To begin, we can represent the observations and the actions of the pilot as follows:

$$\begin{aligned}
\mathcal{B}_{init} &\equiv \{\neg\,\text{alarm}_0, \neg\,\text{buffet}_0, \neg\,\text{stall}_0, \neg\,\text{overspeed}_0\}\\
\\
Obs_1 &\equiv \{\text{buffet}_1, \text{alarm}_1, \text{acceleration}_1\}\\
Obs_2 &\equiv \{\neg\,\text{buffet}_2, \neg\,\text{alarm}_2, \neg\,\text{acceleration}_2,\\
&\quad \text{FD(pull)}_2, \text{V}_\text{z}\uparrow_2\}\\
Obs_3 &\equiv \{\neg\,\text{buffet}_3, \neg\,\text{alarm}_3, \neg\,\text{acceleration}_3,\\
&\quad \text{FD(pull)}_3, \text{V}_\text{z}\uparrow_3\}\\
Obs_4 &\equiv \{\neg\,\text{buffet}_4, \text{alarm}_4, \neg\,\text{acceleration}_4,\\
&\quad \text{FD(pull)}_4, \text{V}_\text{z}\uparrow_4\}\\
\\
\mathcal{T} &\equiv \{a_1 = \text{pull}_1, a_2 = \text{pull}_2, a_3 = \text{push}_3,\\
&\quad a_4 = \text{pull}_4\}
\end{aligned}$$

To understand these observations and actions, it is necessary to model part of the knowledge of the pilots, which corresponds to the following rules $\mathcal{R}$:

$$\begin{aligned}
R_1 &\equiv \text{buffet}_t \Rightarrow \text{stall}_t\\
&\quad \text{vibrations indicate a stall situation}\\
R_2 &\equiv \text{alarm}_t \Rightarrow \text{stall}_t\\
&\quad \text{the stall alarm indicate a stall situation}\\
R_3 &\equiv \text{acceleration}_t \Rightarrow \text{overspeed}_t\\
&\quad \text{an acceleration is an indicator of overspeed}\\
R_4 &\equiv (\text{V}_\text{z}\uparrow_t \wedge \neg\,\text{stall}_t) \Rightarrow \text{overspeed}_t\\
&\quad \text{excluding stall situation, the increase of}\\
&\quad \text{V}_\text{z} \text{ corresponds to an overspeed}\\
R_5 &\equiv \text{overspeed}_t \Rightarrow (\text{pull}_t :: \neg\,\text{overspeed}_{t+1})\\
&\quad \text{pulling the stick solves the overspeed}\\
R_6 &\equiv \text{stall}_t \Rightarrow (\text{push}_t :: \neg\,\text{stall}_{t+1})\\
&\quad \text{pushing the stick solves the stall}\\
R_7 &\equiv \text{FD(pull)}_t \Rightarrow \text{pull}_t\\
&\quad \text{the operator should pull the stick when the}\\
&\quad \text{flight director tells so.}
\end{aligned}$$

Moreover, we add two desires:

$$\mathcal{D} \equiv \{\,\neg\,\text{stall}, \neg\,\text{overspeed}\,\}$$

Finally, to include the basic behavior rules in our system, we add the following permanent rules $\mathcal{C}$:

$$\begin{aligned}
C_1 &\equiv (\text{stall}_t \wedge \text{overspeed}_t) \Rightarrow \bot\\
&\quad \text{stall and overspeed are mutually exclusive;}\\
C_2 &\equiv (\text{push}_t \wedge \text{pull}_t) \Rightarrow \bot\\
&\quad \text{idem for pulling and pushing the stick.}
\end{aligned}$$

### C. Presentation and analysis of the results

The BEA proposes the following factors to explain the behavior of the pilots:

1) "the buffeting, which can be associated in his mind with high speed" (p.186)
2) "It is possible that an attentional selectivity has reduced his ability to perceive the [stall] alarm." (p.188)
3) "Moreover, the presence of the flight director leading to the display of a nose-up attitude [pull on the stick] may have confirmed the [pilots] in the idea that the stall warning was not relevant." (p.187)

The computation of MCSes in our model leads to 903 scenarios (which corresponds to an average branching factor

of 2.34 in our beliefs tree). The analysis of the biases on these scenarios leads us to identify three families:

- The scenarios in which the bias patterns cover all three factors highlighted by the BEA;
- The scenarios where the bias patterns do not correspond (at least, not completely) to the report of the BEA but could explain the crash;
- The absurd scenarios where the MCS do not seem to reflect a plausible behavior.

*a) Scenarios in accordance with the BEA analysis:* In the first family we find for example the following scenario:

| | | |
|---|---|---|
| $MCS(t=1)$ | $\rightarrow$ | $R_1, Obs_1(\text{alarm}_1)$ |
| $MCS(t=2)$ | $\rightarrow$ | $R_5$ |
| $MCS(t=3)$ | $\rightarrow$ | $Obs_3(\text{FD}(\text{pull})_3)$ |
| $MCS(t=4)$ | $\rightarrow$ | $R_2$ |

By ignoring the rule R1 at step 1, the pilot considers that the association "buffet implies stall" is no longer valid. Therefore the pilot can associate the vibration with something else, possibly high speed (assuming that the pilot believes he is in high speed situation). This ignorance of R1 corresponds to the first factor in the BEA's analysis.

The attentional selectivity also mentioned in the BEA report (2) refers to the attention bias we have presented in subsection IV-B and that we also find at step 1: the observation of the alarm $Obs_1(\text{alarm}_1)$ is being ignored while the operator focuses his attention on overspeed.

This mental state scenario also proposes the ignorance at step 2 of rule $R_5$, which states that the action "pull" should have brought the aircraft out of the overspeed situation. The action did not produce the expected effect, according to the observations, which is the reason why the operator must ignore this rule (to maintain the coherence of its belief base). The report of the BEA does not mention such a reasoning but we can nonetheless make a connection with the *commitment* bias presented in subsection IV-C.

Finally, the ignorance of the observation $\text{FD}(\text{pull})_3$ at step 3, *i.e* the failure to follow the instructions given by the flight director, can be interpreted as the pilot realizing that these instructions were wrong and trying to change the strategy. However, at step 4, the stall alarm turns back on and the pilot retains this observation in our scenario (the MCS at t=4 does not contain the alarm observation). To maintain the consistency of the system, the pilot has to ignore rule $R_2$, which connects the alarm and the stall situation: this is our confirmation bias pattern, as introduced in subsection IV-E. This corresponds to not considering the alarm as relevant, as proposed by the BEA in factor (3).

In this scenario, not only are all of the BEA factors retrieved, but we also identify that the *commitment* bias could have played a role with the attention and confirmation biases.

*b) Other scenario matching the analysis by the BEA:* Still in the first family we find the following example:

| | | |
|---|---|---|
| $MCS(t=1)$ | $\rightarrow$ | $R_1, Obs_1(\text{alarm}_1)$ |
| $MCS(t=2)$ | $\rightarrow$ | $Obs_2(\text{V}_\text{z} \uparrow_2)$ |
| $MCS(t=3)$ | $\rightarrow$ | $R_5, Obs_3(\text{FD}(\text{pull})_3)$ |
| $MCS(t=4)$ | $\rightarrow$ | $R_2$ |

In this scenario we find again the attention bias at step 1 and the confirmation bias at step 4. However, rule $R_5$ is not ignored at step 2 (the observation of the vertical speed is ignored instead). This corresponds to an overconfidence from the pilot: he believes at $t = 1$ to be in overspeed and he assumes that the pull action will allow him to get out of this situation. Therefore the pilot does not pay attention to the new information that is inconsistent with his prediction. At $t = 3$, the pilot becomes aware of his mistake and considers that his action is wrong and that the flight director gives bad indications. At $t = 4$ we find an attention bias that does not allow the pilot to consider a stall situation. Thus we find a scenario combining overconfidence, attention bias and confirmation bias which can explain the reasoning of the pilot and comply with the factors identified by the BEA.

*c) Scenario that differs from the BEA analysis:* Consider the following scenario produced by our model as a possible explanation of the behavior of the pilot:

| | | |
|---|---|---|
| $MCS(t=1)$ | $\rightarrow$ | $R_2, Obs_1(\text{buffet}_1)$ |
| $MCS(t=2)$ | $\rightarrow$ | $R_5$ |
| $MCS(t=3)$ | $\rightarrow$ | $Obs_3(\text{FD}(\text{pull})_3)$ |
| $MCS(t=4)$ | $\rightarrow$ | $\emptyset$ |

In this scenario the attention bias is on the vibration and not on the alarm. Moreover, the pilot ignores rule $R2$ and does not connect the stall and the alarm. This ignorance does not correspond to a confirmation bias because no information at $t = 1$ confirms or contradicts the pilot's belief. At first sight, the ignorance of such a rule can seem too irrealistic. However, this possibility is given by the BEA and could be explained by "the low exposure [...] in continuous training (theoretical and practical) to the phenomenon of stalling, to the stall alarm" (p.196). If the pilot does not associate in his mind the stall to the alarm, then we can build a scenario where this ignorance, followed by a commitment bias, does not lead to a confirmation bias. At $t = 4$, the pilot finds himself in a situation where he has no idea what to do. This puzzlement can be seen in the transcription of the cockpit voice recorder: "we have lost control of the plane, we don't understand a thing, we tried everything" (Appendix 1 p.28).

Another scenario in the same family consists in ignoring $\{R_6, Obs_1(\text{acceleration}_1)\}$ at step 1. The pilot does not know the procedure to follow in case of a stall (maybe because of a lack of training).

*d) Scenarios to discard:* Our model also produces some absurd scenarios, such as the one below:

| | | |
|---|---|---|
| $MCS(t=1)$ | $\rightarrow$ | $R_1, R_2, R_3$ |
| $MCS(t=2)$ | $\rightarrow$ | $R_5$ |
| $MCS(t=3)$ | $\rightarrow$ | $R_7$ |
| $MCS(t=4)$ | $\rightarrow$ | $\emptyset$ |

In this scenario, the pilot ignores at step 1 all the rules allowing him to identify a stall or an overspeed. While this is a possible MCS, it seems impossible for a professional pilot: it would mean that the first action was selected randomly.

## VI. Related work

The computer science literature shows only few works on the representation of biases in a decision making process. Most of these works focus on predictive models in a predefined situation such as vaccination campaigns (Voison [15]) or strategic operations (military, diplomatic) (Kulick [16]). The finite state automaton model of Voison and the blackbox model of Kulick do not offer the possibility of adapting to decision-making situations outside their design and do not allow to implement other biases. In contrast, we aim at a diagnosis model that can explain a behavior in various situations with several biases. On the other hand, the BDI based model of Arnaud and al. [17] allows the implementation of biases with a function that increases or decreases the probability of a belief for each bias. As we stated previously, 151 biases are present in the literature, which means 151 possible functions without having the guaranty that they do not overlap, as there is no consensus to date on the taxonomy of biases [18]. This is why we differ, by basing our model on a diagnosis approach in order to catch as many biases as possible and not to limit the explanation of a reasoning to a single bias. Finally we can mention Dutilh Novaes and al. [19] who are the closest to our work by using an operator of minimal belief revision to predict the behavior of the agent in tasks related to a belief bias. Although our model is based on the same revision operator, we propose an explanatory model for identifying several biases.

## VII. Conclusion & Perspectives

Our model uses belief revision-based diagnosis and cognitive biases to identify a set of rational mental state scenarios that explain an irrational behavior. It proposes, for each scenario, a trace of events and belief revisions that correspond to the different biases. This model has the advantage of relying on a strong theoretical grounding of over 35 years of research [20], facilitating the addition of future extensions.

One limitation of our model is the difficulty to differentiate closely related biases or to decide which bias is involved among several possibilities. We are working on the identification of logical patterns common to several biases (e.g the preference of old beliefs to new information that we find in confirmation bias and anchor bias). It will allow us to present to the user a synthetic vision of the different possible biases.

We also intend to extend our set of bias patterns and to implement a systematic search algorithm that will provide a clear presentation of the selected biases. Indeed, on the Rio-Paris example, the analysis of each belief revision track is done manually, based on the patterns that correspond to the BEA analysis. Our goal is to create a taxonomy of biases, inspired by the taxonomies proposed in the literature in human sciences, that will draw a partitioning of MCSes.

Moreover we would like to validate the outputs of our model with the domain experts, in particular, those that differs of the BEA analysis.

Finally, we would like to extend our model to include emotions. Several works (see for example [10] on the attention bias), report the importance of emotions in the biases. A pilot with a traumatic experience of running out of fuel will try to avoid at any cost this kind of situation (and thus to take other risks). We wish to rely on affective models using the BDI logic, like in [21], [22]. Our model provides a solid basis for taking into account a wide range of irrationality factors in decision-making situations while addressing the limitations of current cognitive bias models.

## References


[1] BEA, "Bea f-cp090601," tech. rep., Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, 2012.

[2] P. Gärdenfors and H. Rott, *Belief Revision*, vol. 4, pp. 35–132. 04 1995.

[3] R. Reiter, "A theory of diagnosis from first principles," *Artificial Intelligence*, vol. 32, no. 1, pp. 57 – 95, 1987.

[4] R. Wassermann, "An algorithm for belief revision," tech. rep., 2000.

[5] C. E. Alchourrón, P. Gärdenfors, and D. Makinson, "On the logic of theory change: Partial meet contraction and revision functions," *The journal of symbolic logic*, vol. 50, no. 2, pp. 510–530, 1985.

[6] M. H. Liffiton and K. A. Sakallah, "Algorithms for computing minimal unsatisfiable subsets of constraints," *Journal of Automated Reasoning*, vol. 40, no. 1, pp. 1–33, 2008.

[7] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[8] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 2, pp. 1413–1432, 2020.

[9] A. Murata, T. Nakamura, and W. Karwowski, "Influence of cognitive biases in distorting decision making and leading to critical unfavorable incidents," *Safety*, vol. 1, no. 1, pp. 44–58, 2015.

[10] C. MacLeod, A. Mathews, and P. Tata, "Attentional bias in emotional disorders.," *Journal of abnormal psychology*, vol. 95, no. 1, p. 15, 1986.

[11] B. M. Staw, *The escalation of commitment: An update and appraisal*, p. 191–215. Cambridge Series on Judgment and Decision Making, Cambridge University Press, 1996.

[12] D. A. Moore and P. J. Healy, "The trouble with overconfidence.," *Psychological review*, vol. 115, no. 2, p. 502, 2008.

[13] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[14] S. Conversy and al., "L'accident du vol AF447 Rio-Paris, un cas d'étude pour la recherche en IHM," in *IHM'14, 26e conférence francophone sur l'Interaction Homme-Machine*, pp. 60–69, ACM, Oct. 2014.

[15] M. Voinson, S. Billiard, and A. Alvergne, "Beyond rational decision-making: modelling the influence of cognitive biases on the dynamics of vaccination coverage," *PloS one*, vol. 10, no. 11, 2015.

[16] J. Kulick and P. K. Davis, "Modeling adversaries and related cognitive biases," *Modeling Adversaries and Related Cognitive Biases*, 2003.

[17] M. Arnaud, C. Adam, and J. Dugdale, "The role of cognitive biases in reactions to bushfires," in *ISCRAM*, (Albi, France), May 2017.

[18] A. Ceschi, A. Costantini, R. Sartori, J. Weller, and A. Di Fabio, "Dimensions of decision-making: An evidence-based classification of heuristics and biases," *Personality and Individual Differences*, vol. 146, pp. 188–200, 2019.

[19] C. Dutilh Novaes and H. Veluwenkamp, "Reasoning biases, non-monotonic logics and belief revision," *Theoria*, vol. 83, 12 2016.

[20] E. Fermé and S. O. Hansson, "Agm 25 years: Twenty-five years of research in belief change," *Journal of Philosophical Logic*, vol. 40, pp. 295–331, 04 2011.

[21] M. Dastani and E. Lorini, "A logic of emotions: from appraisal to coping.," in *AAMAS*, pp. 1133–1140, 2012.

[22] C. Adam, A. Herzig, and D. Longin, "A logical formalization of the occ theory of emotions," *Synthese*, vol. 168, no. 2, pp. 201–248, 2009.