HAL
open science

Predicting Human Operator's Decisions Based on Prospect Theory

Paulo Eduardo Ubaldino de Souza, Caroline Ponzoni Carvalho Chanel, Melody Maillez, Frédéric Dehais

To cite this version:

Paulo Eduardo Ubaldino de Souza, Caroline Ponzoni Carvalho Chanel, Melody Maillez, Frédéric Dehais. Predicting Human Operator's Decisions Based on Prospect Theory. Interacting with Computers, 2020, pp.1-16. 10.1093/iwcomp/iwaa016 . hal-02923113

HAL Id: hal-02923113

https://hal.science/hal-02923113v1

Submitted on 26 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive Toulouse Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of some Toulouse researchers and makes it freely available over the web where possible.

This is an author's version published in: https://oatao.univ-toulouse.fr/26587

Official URL: https://doi.org/10.1093/iwcomp/iwaa016

To cite this version :

Ubaldino de Souza, Paulo Eduardo and Ponzoni Carvalho Chanel, Caroline and Maillez, Melody and Dehais, Frédéric Predicting Human Operator's Decisions Based on Prospect Theory. (2020) Interacting with Computers. 1-16. ISSN 0953-5438

Any correspondence concerning this service should be sent to the repository administrator: tech-oatao@listes-diff.inp-toulouse.fr

Predicting Human Operator's Decisions Based on Prospect Theory

Paulo E. U. de Souza, Caroline P. C. Chanel*,
Melody Mailliez and Frédéric Dehais

ISAE-SUPAERO, Université de Toulouse,
10 Av. Edouard Belin 31400 Toulouse, France
**corresponding author: caroline.chanel@isae-supearo.fr*

Abstract

The aim of this work is to predict human operator's decisions in a specific operational context, such as a cooperative human-robots mission, by approximating her utility function based on Prospect Theory. To this aim, a within-subject experiment was designed in which the human operator has to decide with limited time and incomplete information. This experiment also involved a framing effect paradigm, a typical cognitive bias causing people to react differently depending on the context. Such an experiment allowed to acquire data concerning the human operator's decisions in two different mission scenarios: search and rescue and Mars rock sampling. The framing was manipulated (e.g. positive vs. negative) and the probability of the outcomes causing people to react differently depending on the context. Statistical results observed for this experiment supported the hypothesis that the way the problem was presented (positively or negatively framed) and the emotional commitment affected the human operator's decisions. Thus, based on the collected data, the present work is willed to propose: (i) a formal approximation of the human operator's utility function founded on the Prospect Theory; and (ii) a model used to predict the human operator's decisions based on the economics approach of multi-dimensional consumption bundle and Prospect Theory. The obtained results, in terms of utility function fit and prediction accuracy, are promising and show that similar modeling and prediction method should be taken into account when an intelligent cybernetic system drives human-robots interaction. The advantage of predicting the human operator's decision, in this operational context, is to anticipate her decision, given the way a question is framed to the human operator. Such a predictor lays the foundation for the development of a decision-making system capable of choosing how to present the information to the operator while expecting to align her decision with the given operational guideline.

Keywords: *Human-computer interaction (HCI); Human-robot(s) Interaction (HRI), Model-based Predictor Design; Prospect Theory, Cognitive Bias; Framing Effect.*
Handling Editor: *Russell Beale*
Received 26 September 2018; Revised 31 October 2019; Accepted 19 May 2020

Research Highlights

The research contributions of this work can be declined in the following topics:

- The presentation of an human-drones interaction experiment performed to collect data, in which the human operator has to decide with limited time and incomplete information;
- A statistical study to verify the relationship between the Human Operator's decision and mission context variables;
- The proposition of an Human Operator's utility function based on Prospect Theory, for which the parameters were approximated from experimental data including the significant results of the statistical study;
- The proposition of a model-based predictor, inspired by the economics approach of multi-dimensional consumption bundle and Prospect Theory;
- The evaluation of the proposed decision model, the which predict's de Human Operator's decision depending on the mission context and on the framing presented to the operator.

Interacting with Computers, 2020, Published by Oxford University Press on behalf of The British Computer Society
DOI: 10.1093/iwcomp/iwaa016

1 Introduction

In recent years, there has been a growing interests about the use of totally autonomous robots to replace humans in a variety of dirty, dull and dangerous missions. Recent technical progress as well as advanced artificial intelligence have allowed to design vehicles and robots with high level of control and decisional autonomy (Matarić et al., 2003; Timotheou and Loukas, 2009; Murphy et al., 2008; Suarez and Murphy, 2011; Xue et al., 2011; Kolling et al., 2016). However, in complex missions, these autonomous systems still require to be supervised by humans to ensure the smooth conduct of the mission (Schurr et al., 2009). These latter are expected to take over during unforeseen situation (de Winter and Dodou, 2014) or when ethical concerns are at stake (Belloni et al., 2014). These issues raise the importance of human-machine teaming as the next challenge to optimize the efficiency and the safety of operations. Indeed, careless design of user interface, inadequate task allocation and poor authority implementation between human and artificial agents can dramatically impair human operators performance (Dehais et al., 2015, 2005) to an extent that they can persist in irrational decision making (Dehais et al., 2012, 2019).

In this context, mixed-initiative interaction provides a relevant framework as it considers that the agents' (human and robot) abilities are complementary and are likely to provide better performance when joined efficiently than when used separately. In particular, in the human-robots interaction community (HRI), *mixed-initiative* implies that humans or artificial agents can seize initiative from each other by themselves (Jiang and Arkin, 2015). The implementation of mixed-initiative interaction driving systems presuppose to develop algorithms dedicated to learn how human make decisions in order to optimize teaming with artificial agents (Guo et al., 2018; Gombolay et al., 2017).

A first step toward the implementation of such algorithms is to consider theories related to human decision making. These theories that can be roughly separated in two classes: prescriptive and descriptive approaches (Baron, 2007; Kahneman, 2011; Bago and De Neys, 2017). The *prescriptive approach* explores how people should make optimal decisions. It typically assumes ideal circumstances, as for instance, complete information, awareness of all options, abundance of time to decide to model the best and rational path such that a person comes to the most suitable decision (Todd and Gigerenzer, 2000). Hence, it is assumed that decision makers have an *utility function*, and they are always trying to maximize their utility from a stable set of preferences (Suhonen, 2007). In fact, some important contributions to descriptive theories of thinking are not obtained by observing people's thinking but from attempts to make computers think (Baron, 2007).

The *descriptive approach* focuses on how humans actually make decisions in complex and uncertain real life situations. For instance, Kahneman (2011) postulated that humans enjoy two opposite mode of thinking that are analytical an intuitive. On one hand, *analytical thinking* is *slow and effortful* but is logical, flexible and generally yield to allows optimal and effective conclusions. Human beings use analytical thinking when facing novel situations especially it they are not time constrained. On the other hand, *intuitive thinking* is fast, automatic, often unconscious, requires few cognitive resources and is used during most of our routine operations. By generalizing circumstances, it allows us to reduce the complexity of a situation, recognize patterns (real or perceived) and make decisions quickly according to past experiences or the logic of those recognized patterns. However, while this mode of thinking is exceptionally efficient and accurate, it is biased and prone to errors especially under emotional settings Biswas and Murray (2017); Robinette et al. (2016); Kahneman (2011).

For instance, Kahneman (2011) demonstrated that *Framing Effect (FE)* can bias intuitive mode of thinking when real-life problems are presented in positive (i.e. gain) or in negative (i.e. loss). *Framing Effect (FE)* theory is a strong and powerful finding (see (Steiger and Kühberger, 2018) for a recent meta-analysis) explaining that human beings tend to be risk-averse when positive frames are presented but risk-seeking when a negative frame are presented. Researches have shown that losses evoke stronger negative feelings than gains and choices are not reality-bound because intuitive thinking is not bound to reality (Baron, 2007). In other words, the frame significantly affects how people infer meaning and hence understands the situation. Unless there is a clear reason to do otherwise, most of people passively accept decision situations as they are framed (Kahneman, 2011), because reframing is arduous and analytical thinking is typically lazy. Levin et al. (1998) postulated the existence of three types of framing effects: (1) Risk Choice Framing (RCF) (Tversky and Kahneman, 1981), which involves options differing in level of risk and described in different ways; (2) Attribute Framing (AF), which affects the evaluation of the characteristics of an event or object; and (3) Goal Framing (GF), which affects the persuasiveness of a communication. Note that, *Attribute Framing (AF)* seams to be the simplest case of framing, where only a single attribute is the subject of the framing manipulation and the evaluation can be measured by choices between yes or no.

Overall, these findings indicate that modelling human behavior is not a straightforward task and can't be reduced in terms of simple computation of expected utilities. There is a need to better understand human decisions. In particular when interacting with robots under uncertain settings. Thus, to collect data of human behavior in such a context would allow to design models and algorithms to formalize decision making and to improve humman-machine teaming on the

long run.

The present set of work is part of a wider research in a "robust mixed-initiative multi-agent planning, control, and execution framework", in which the robots have to infer an *human utility* in order to maximize the joint system performance. In this work, the "system" includes the human agent as part of the team, and for that it should take into account the capacities and restrictions of each agent (human operator and robots). In this context, *robust* means that the performance of the system is satisfactory even when reality differs from assumptions. The aim is, in future work, to consider human cognitive biases (1) to adjust the robot team utility function and (2) to adapt dynamically human-drones interaction.

In this regard, a within-subject experiment was designed to collect data related with the human behavior when interacting with robots under uncertain settings in two different contexts. The two different scenarios considered were: (i) helping victims of an earthquake and (ii) sampling rocks on Mars. Twenty participants have performed this experiment, and, the presence of a *Framing Effect (FE)*, a cognitive bias, in which one's reaction differs depending on the way the problem is presented (Kahneman, 2011) was observed. The results supported the hypothesis that the way the problem was presented (positively or negatively framed), and the emotional commitment, statistically affected the HO's decisions. Then, based on the data collected among the 20 participants, the present work proposes:

- in a first step, a HO *Prospect theory (PT)* gain-loss function approximation (Kahneman and Tversky, 1979a) to describe the HO's utility function. In the scenarios considered, aerial robots should search and locate potential targets, while HO decides, with incomplete information and limited time, when and where an aerial robot had to take an action. In this sense, the use of PT is justified since it models how intuitive thinking influences people's immediate reaction to a risk or gamble they are facing;
- then, in a second step, this utility function becomes part of a decisional model used to predict the HO's decisions, based on the economics approach of multi-dimensional consumption bundle (Kőszegi and Rabin, 2006) and PT, without any simplifying assumption.

Results concerning utility function fit and prediction accuracy suggest that such a methodology and prediction model could be used to infer human operator's decisions in a given context.

This paper is organized as follows: The first section presents the *Prospect Theory (PT)*. Section 3 details the experiment that allowed to collect data to implement the *human utility* based on PT. Section 4 present the model used to predict the HO's decisions. Some discussions about the results are presented in Section 5. Finally conclusions and future work are discussed in Section 6.

2 Prospect Theory

Hence, research is carried out based on the *Prospect theory* (PT) (Kahneman and Tversky, 1979a; Tversky and Kahneman, 1992) to consider human's cognitive biases, in particular the *framing effect* (FE), in order to learn a function that fits the HO utility, and helps to predict her decision.

Kahneman and Tversky (Kahneman and Tversky, 1979a; Tversky and Kahneman, 1992) formulated the *(Cumulative) Prospect Theory* (PT), which shows the way mental processes affect human decisions. Mainly, how intuitive thinking influences people's immediate reaction to a risk or gamble they are facing.

PT describes the decision processes in two stages: editing and evaluation (Wakker, 2010; Barberis, 2013). In the first stage – *editing* – , decision outcomes are intuitively ordered according to a certain heuristic, which sets a reference point and then considers lesser outcomes as losses and greater ones as gains. In the second stage – *evaluation* –, people behave as if they would compute an expected utility, based on the potential outcomes and their respective probabilities, and then choose the alternative which has a higher utility. The PT expected utility function (Kahneman and Tversky, 1979a), is recalled here as:

$$\mathbb{E}[u] = \sum_{i=1}^{n} pv(x_i)w(p_i) \tag{1}$$

where, $\mathbb{E}[u]$ is the expected utility of the outcomes, pv is a value function that assigns a personal value to outcomes $\{x_i ... x_n\}$, p_i is the respective probability of an x_i, and $w(\cdot)$ is a subjective *probability weighting* function. Kahneman and Tversky (Kahneman and Tversky, 1979a; Tversky and Kahneman, 1992) emphasize that this transformed probability function does not represent erroneous beliefs, rather, they are decision weights. $w(\cdot)$ is a strictly increasing function that satisfies $w(0) = 0$ and $w(1) = 1$ and that may differ between gains and losses (Kőszegi and Rabin, 2006).

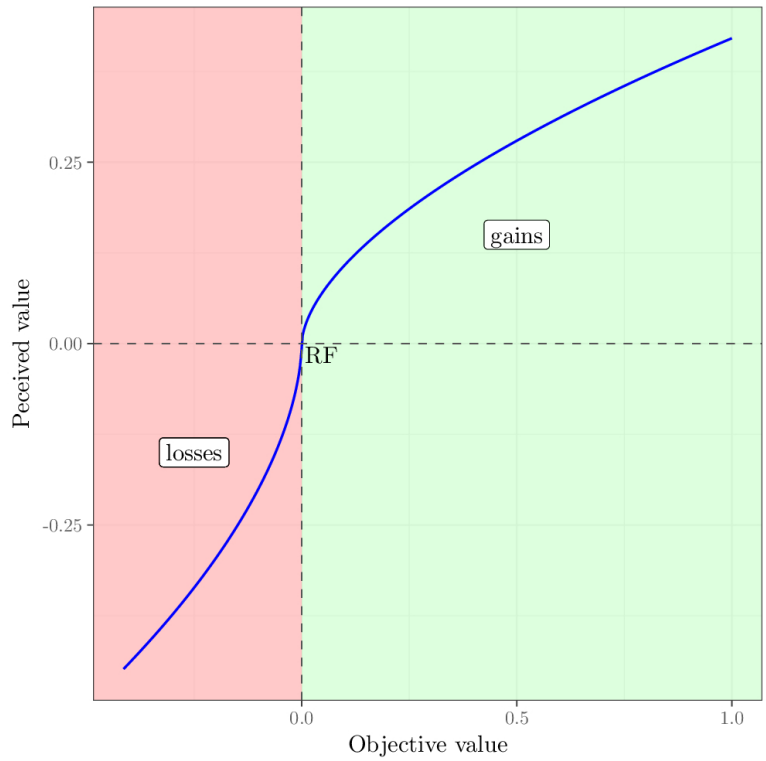

Figure 1: Prospect Theory value function.

The value function $pv(\cdot)$, defined in Equation 2 and shown in Fig. 1, passes through the reference point, is continuous for all objective values x, strictly increasing s-shaped and asymmetrical, leading people to be risk-averse for gains and risk-seeking for losses and, also, showing that losses hurt more than gains feel good. This *loss aversion* is defined by the λ constant factor (see Eq. 2).

$$pv(x) = \begin{cases} x^{\alpha} & x > 0 \\ -\lambda(-x)^{\beta} & x \leq 0 \end{cases} \tag{2}$$

This formulation (Eq. 2) illustrates three elements of PT (Kahneman and Tversky, 1979a) and corresponds to Kahneman and Tversky's explicit or implicit assumptions about their value function:

- *Reference dependence* - people derive utility from gains and losses, measured relative to some reference point, rather than from absolute levels of wealth.
- *Loss aversion* - loss aversion is generated by making the value function steeper, modeled by the λ constant, in the region of losses than in the region of gains. If $y > x > 0$, then $pv(y) + pv(-y) < pv(x) + pv(-x)$.
- *Diminishing sensitivity* - the value function is concave ($pv''(x) \leq 0$ for $x > 0$) in the region of gains but convex ($pv''(x) \geq 0$ for $x < 0$) in the region of losses. It is modeled by the constants α and β. The concavity over gains captures the finding that people tend to be risk averse over probability gains. However, people also tend to be risk seeking over losses.

It is important to notice that the *probability weighting* function, the forth PT element, (see Eq. (1)) models the fact that people does not weight outcomes by their objective probabilities p_i but rather by transformed probabilities or decision weights $w(p_i)$ (Tversky and Kahneman, 1992). In this sense, a delicate issue is how to approximate a such personal weighting function.

Recently, several researchers have used PT to explain the decision-making process. For instance, Nadendla et al. (2016) argues that, for hypothesis of testing, an human agent decision can be model by PT. In Zhang (2016), an emotion driven behavior selection mechanism based on the PT's Value Function is suggested in order to understand the autonomous behavior of artificial life. Ren and her colleagues (Ren et al., 2016) propose a method to deal with the emergency decision making based on PT.

Moreover, Kőszegi and Rabin (2006) propose a formal framework for applying PT in economics. They argue that their proposal is both disciplined and applicable to different contexts. The idea is that the reference point people use to compute gains and losses is fully determined by their expectations (instead of the status quo). In particular, they propose that people derives utility from the difference between consumption and expected consumption, for instance a salary of $\$50,000$ to an employee who expected $\$60,000$ will not be assessed as a gain relative to status-quo wealth, but rather as a loss relative to expectations of wealth. They also assume that expectations are rational, i.e. they match the distribution of outcomes that people will face if they follow the plan of action that is optimal, knowing their expectations. The conclusion is drawn as a person's utility depends on her multi-dimensional consumption bundle c and also on a reference bundle r, combining classical consumption utility with reference dependence utility by assuming people care about both. For instance, they do not just react to the sensation of gaining or losing a mug, but they also care whether they have a mug to drink from. Thus, this *personal Utility (U)* is given by:

$$U(c|r) = m(c) + n(c|r) \tag{3}$$

where, $m(c)$ is an intrinsic "consumption utility" (typically stressed in economics) that corresponds to the personal outcome-based utility, and $n(c|r)$ is a *gain-loss utility*, that should be in accordance with PT, given by:

$$n(c|r) = \mu(m(c) + m(r)) \tag{4}$$

Their model allows for both stochastic outcomes and stochastic reference points, and assumes that a stochastic outcome is evaluated according to its expected utility. For instance, if c is drawn according to the probability measure F, the person's utility is given by:

$$U(F|r) = \int u(c|r)dF(c) \tag{5}$$

However, considering the gain-loss utility $n(c|r)$, they impose some simplifying assumptions, like linear utility for gains and losses and no probability weighting, which differs from the proposed *gain–loss utility* presented in Equation (1).

In summary, the work proposed in the following explores the PT in order to model the human's utility based on data collected among 20 subjects during an experiment (described hereafter), and a decisional model used to predict the HO's decisions, based on the economics approach of multi-dimensional consumption bundle (Kőszegi and Rabin, 2006) and the PT, however, without any simplifying assumption.

3 Experiment for data acquisition

As in this work our first aim was to observe the utility function for gains and losses and the FE influence over the decision taken by the human operator (HO), two different scenarios were considered in an experiment: (1) helping victims of an earthquake and (2) collecting rocks on Mars. Recalling, such a function approximation is necessary to construct the decisional model useful to predict HO's decisions.

It is important to note that in this experiment the same graphical user interface (see Figure 2) was used for both scenarios. The system selected one of them at the beginning randomly, in other words, there are no visual differences between the scenarios, only the back-story was different. In the *Earthquake scenario* the idea was to help eight known victims trapped beneath the rubble. HO (subject) should use three drones to localize the victims and to deliver eight available first-aid kits. A guideline was presented to find and deliver a first-aid kit to the maximum number of victims within a certain time period. In the *Mars rock sampling scenario* HO had three drones to localize and collect eight different types of rocks and return then to Earth in a capsule with eight containers. The guideline was to find and collect the maximum number of "good" rocks before the time was up.

3.1 Participants

Twenty volunteers (34.81% female, mean age: 30.73, sd: 7.54), all graduate students, participated in the experiment. They were, unknown to them, randomly split into two groups (one for each scenario) following a within-subject protocol. They were not rewarded for participation. We note that only 45% of the participants made the experiment in their first language (French or English), however statistical analysis did not show any significant correlation with this variable.

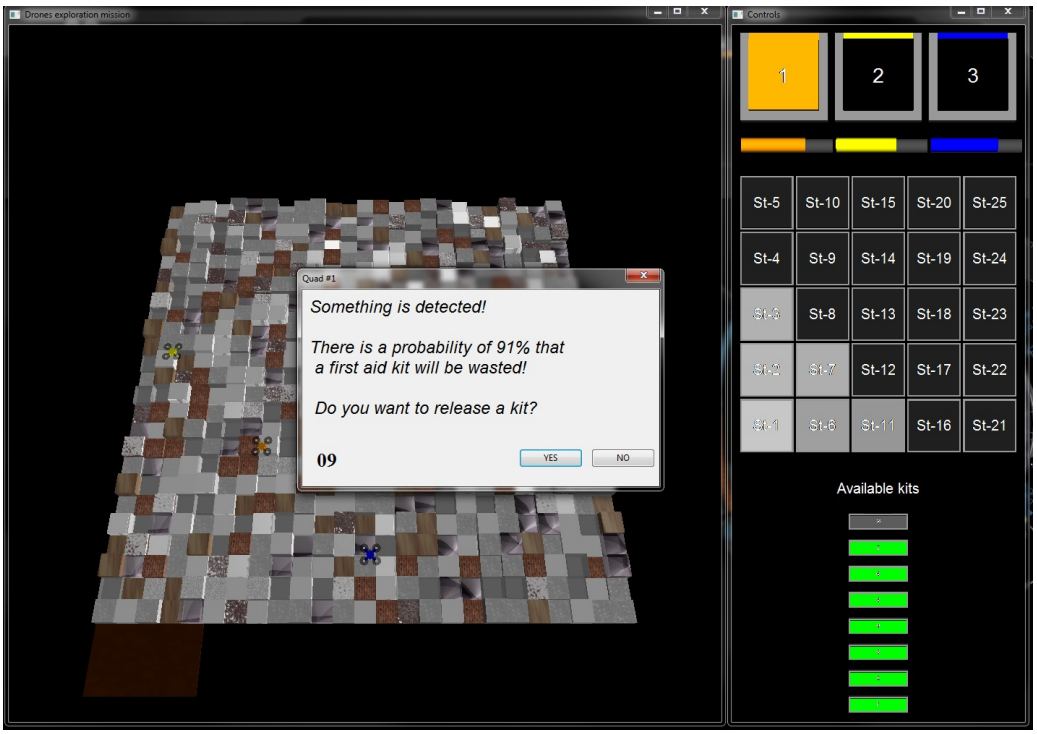

Figure 2: The operator's graphical interface common to both scenarios (developed in Python 2.7.11).

3.2 Experimental protocol

Each participant, after a training mission, randomly executed 10 missions (repetitions) for a given scenario. Each mission had a duration of approximately three minutes. During the evolution of a given mission scenario, when the drones found something, 10 different pre-formulated sentences, relating to the *Text Framing* and to the *Probability*, were randomly presented and the operator was requested to decide.

The *Text Framing* was the way of presenting the questions (positively or negatively). In the earthquake scenario, for the *Positive frame* a sentence was like: *There is* 60% *of chance that the kit will be* ***useful***; and, for the *Negative frame* it could be: *There is* 40% *of chance that the kit will be* ***wasted***. In the case of the Mars rock sampling scenario, for the *Positive frame* a sentence was such as: *There is* 60% *of chance of being a* ***'good' rock***; and: *There is* 40% *of chance of being a* ***'bad' rock***, in other case. Note the *Text Framing* has the form of an *Attribute Framing* (AF), the simplest case of framing, where only a single attribute is the subject of the framing manipulation (e.g. the probability) and the evaluation can be measured by choices between yes or no.

For the *Probability*, we could be interested in the values close to 50% where the attribute framing is more effective. However, in order to evaluate the *power* of the framing, the values presented were uniformly selected from 0.01 to 0.99. Then, this range of probability values was discretized into four probability levels, e.g. *low*, *middle-low*, *middle-high* and *high*. And, each probability level has two sentences: one with a positive frame and the other with a negative one.

The participants had 10 seconds to decide between *YES*, i.e, take a *positive action* (release a first-aid kit or collect a rock), and *NO*. At the end of the alloted time, if the human operator (e.g. participant) has not selected an answer for the question asked, the drone who asked should consider the human operator's decision as a *NO*. The only task of the human operator was to answer the questions asked by the drones.

Note that it was not possible for the participants to know the real result of the mission, i.e., the participant could not know how many victims were helped or *good* rocks were collected during the experiment.

3.3 Statistical results based on experimental data

We analyzed 1982 observations from 20 participants [1]. Because we took multiple measures per subject (within-subject protocol) and the dependent variable (HO's decision) had a binomial distribution, a *Generalized Linear Mixed Model - GLMM* was used (Agresti and Kateri, 2011; Bates et al., 2015).

[1] Anonymized data is available at:
`https://personnel.isae-supaero.fr/IMG/all-more.csv`

In this study we are interested in the relationship between HO's decision ($\boldsymbol{OD}$) and the main explanatory variables:

- the Scenario ($\boldsymbol{S}$);
- the Text framing ($\boldsymbol{TF}$):
- the Probability ($\boldsymbol{P}$) that a kit would be useful or not (earthquake scenario) or the target would be a *good* rock or not (Mars scenario);
- and, the number of used Assets ($\boldsymbol{A}$), that would be first-aid kits (earthquake scenario) or containers (Mars scenario).

Note that, $\boldsymbol{A}$ was used to determine the HO *Reference point* (according to PT, see Sec. 4) and when she considered the action either as a gain or as a loss. The following equation describes the relation between operator's decision ($\boldsymbol{OD}$) and the explanatory variables:

$$\boldsymbol{OD} \sim \boldsymbol{S} + \boldsymbol{TF} + \boldsymbol{P} + \boldsymbol{A} + (1|\boldsymbol{ID}) + (1|\boldsymbol{Seq}) + \epsilon \tag{6}$$

It is important to notice that we started the statistical analysis with a model with all fixed effects available and dropped one by one until all unnecessary terms were removed, for instance: age, gender and language (native or not).

The random factors that were not possible to control experimentally, were unpacked in two different variables: $\boldsymbol{ID}$ and $\boldsymbol{Seq}$. The first one refers to the assumption of a different intercept for each subject and the second one refers to the sequence of the missions, which were shuffled for each subject. All the others "stochastic" differences are retained in terms of error ϵ.

Table 1 shows the estimated coefficients and errors of the GLMM. Here, the positive value of an estimated coefficient denotes that the condition increases the preference in saying "YES". Here, the Earthquake scenario led the participants to say more "YES" than the Mars scenario. The same is observed in the Positive frame condition. For Probability and used Assets, the coefficients denote that increasing the probability value or the number of used assets also increase the willingness to say "YES". The intercept is the predicted value of decision when all the independent variables are 0.

Table 1: GLMM summary

	Dependent variable:
	Decision (OD)
Earthquake (S)	0.323^{*}
	(0.192)
Positive frame (TF)	0.252^{**}
	(0.127)
Probability (P)	0.731^{***}
	(0.279)
Asset (A)	0.311^{***}
	(0.042)
Intercept	-0.938^{***}
	(0.235)
Log Likelihood	−735.662
Akaike Inf. Crit.	1,485.324
Bayesian Inf. Crit.	1,520.955
Note:	*p<0.1; **p<0.05; ***p<0.01

4 Proposed PT Model

As a first step, based on the collected data, PT is explored in order to formally describe the HO's utility function. Later, the authors proposed a model, which predicts the HO's decision given the explanatory variables (scenario, text framing, probability and the number of assets). This decisional model is based on a decisional mathematical criterion, and should be used, in future work, to decide how to frame a question to HO, in order to, at least, maximize the chances to induce a desired decision from her in a given operational scenario.

In the following is presented our estimation for the intrinsic *consumption utility*, $m(c)$, and the gain-loss function, $n(c|r)$ (see Equation 3). This step helps in determining the personal utility function $U(\cdot)$ as proposed by Kőszegi and Rabin, but without any simplification assumption - contrary to what has been proposed in (Kőszegi and Rabin, 2006) as linear utility for gains and losses and any probability weighting function.

4.1 Approximating the intrinsic consumption utility

According to our hypothesis, the HO had two personal *goods* in the previous experiment: the *belief that her action could result in a good job* (helping victims or collecting rocks - $h \in \{0, 1\}$) and the perceived ownership value of an asset (available kits/containers - $pc_a \in \mathbb{R}$) at a given moment. Hence, she had a bi-dimensional consumption bundle $c = (h, pc_a)$. Note that, HO faced conflicted emotions while deciding: if she said *YES*, she got the *satisfaction* (gain) of doing a good action against the probability of loosing a precious asset. Else, by saying *NO*, she could save the asset for a better future opportunity, but with the risk of leaving behind a victim or a wanted rock.

In this cost-benefit dilemma, supposing that HO wanted to do a good work, for instance by doing a *positive action* (saying *YES*) in function of her consumption bundles, the function $m(c)$ could be, in this application case, as:

$$\begin{aligned} m(c) &= c(h, pc_a, p) \\ &= \begin{cases} (h^+ - pc_a) \cdot p, \text{for a positive action} \\ (pc_a - h^-) \cdot (1 - p), \text{for a negative action} \end{cases} \end{aligned} \tag{7}$$

where p is a given objective probability value, or the presented one, for instance.

4.2 Approximating the gain-loss utility

Following Equations (3) and (4) the term $n(c|r) = \mu(m(c) + m(r))$ represents the *gain–loss utility*, that is based on PT (cf. Section 2) in this work. For this propose, one should define the weighting $w(.)$ and the perceived value $pv(.)$ functions following Equation (1).

4.2.1 Definition of the probability weighting function - $w(\cdot)$

Since the dependent variable OD (see Eq. 6) was represented by "1" and "0" (*YES* or *NO* resp.) instead of cardinal numbers, a *Binomial Logistic Regression* was used to describe the average preference of the participants as a function of the probability values. Figure 3 shows the probability of saying *YES* versus the objective probability, according to the framing used, with the logistic regression curve fitted to the data. The analysis gives the result presented in Table 2.

From the point of view of the decisions, i.e. OD as a function of the objective probability P, the result indicates that the probability of saying *YES* is significantly associated with the probability of a kit being useful (in the earthquake scenario) or a rock being a *good one* (rock sample Mars scenario). Here, it is easy to observe the *framing effect* with a significant difference between positive and negative framing. Also, by comparing the results of different scenarios, the influence of *emotional commitment* on the subject's decisions is observed. These statistical effects, related with the emotional commitment, were also observed in our previous work (Souza et al., 2016), but for only 14 subjects.

Interestingly, each curve presented in Figure 3 is strictly increasing function that satisfies $f(0) = 0$ and $f(1) = 1$, which is required by PT. From this, it is possible to derivate the *subjective probability weighting function* $w(\cdot)$, as:

$$w(p) = \frac{1}{1 + e^{-(I+bp)}} \tag{8}$$

where, I is the *Interceptor*, b is the estimated coefficient and p is the objective probability.

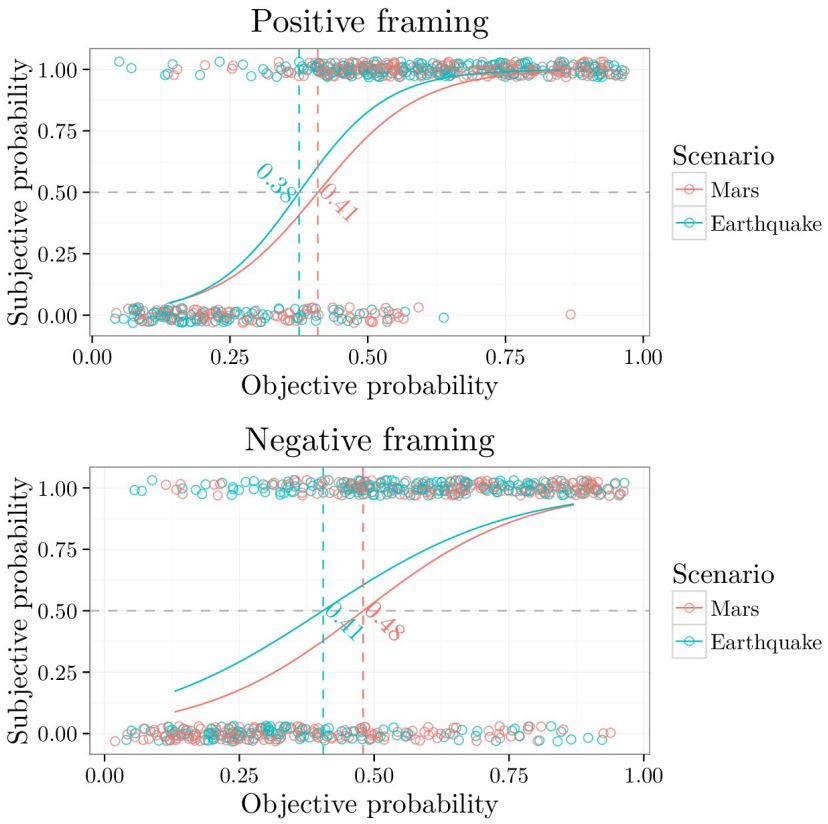

Figure 3: Probability of saying "YES" according to the objective probability.

4.2.2 Approximating personal value function - $pv(\cdot)$

From Equations (1) and (2), one should approximate the *personal value function* $pv(\cdot)$ (according to PT presented in Sec. 2).

In Figure 4, HO's average preference is evaluated against the number of assets used. A *Binomial Logistic Regression* is used again to describe this relationship. It suggests that, in the beginning, the participants had an *endowment effect* for the assets (they became *owners*) which decreased their willingness to use them. This attachment effect made them consider the use of an asset as a *loss* and only accepted to give up of it by a *high price* (high probability value). However, at a certain point they changed their mind and started to act as "sellers" (which do not assess *sales* as loss of inventory but as a gain of money). After the *reference point*, the graph suggests that they want to use the assets as much as possible.

Such a hypothesis was also observed in our previous work, where a survey was conducted after each mission and it

Table 2: Logistic regression analysis

Framing	Scenario	variable	Coefficient	error
Positive	earthquake	Intercept	-5.924^{***}	0.983
		Probability	16.437^{***}	2.219
	rock sampling	Intercept	-7.403^{***}	1.311
		Probability	18.373^{***}	2.772
Negative	earthquake	Intercept	-3.442^{***}	1.178
		Probability	9.103^{***}	2.256
	rock sampling	Intercept	-5.287^{***}	1.269
		Probability	11.609^{***}	2.512

Note: *p<0.1; **p<0.05; ***p<0.01

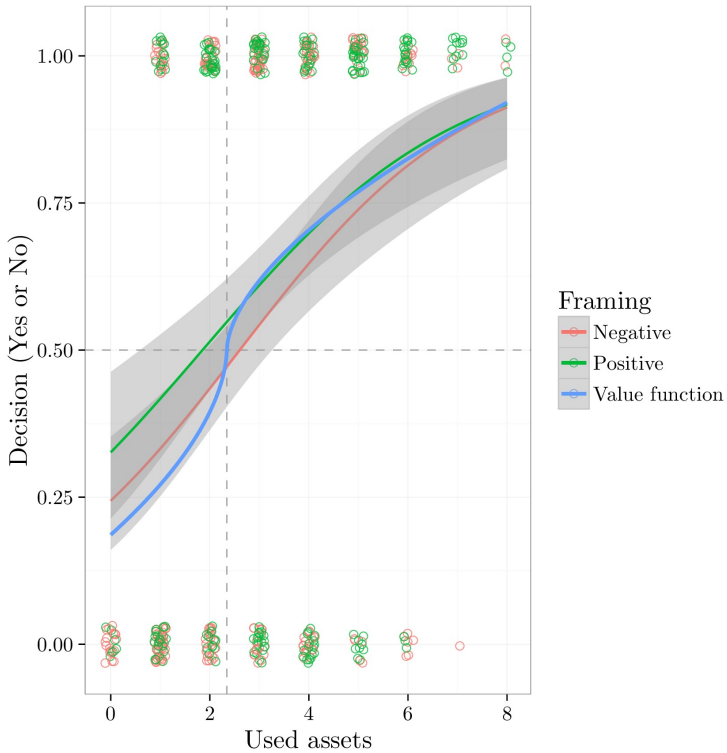

Figure 4: Probability of saying "YES" according to the number of used assets. The dark blue line is the approximated personal value function of Equation (2).

showed that the participants were more satisfied when they used more assets.

Thereby, one can assume the presence of a *reference point* RF across which the preference of saying *YES* becomes greater than the preference of saying *NO*. Thus, a curve that satisfies a *PT value function* $pv(\cdot)$ (see Eq. (1)) could be fitted over the positive framing curve in the *gain* region and over the negative framing curve in the *loss* region.

Without loss of generality, it is reasonable to assume that the *perceived ownership value* $pc_a(\cdot)$ (see Eq. (7)) of an asset in a given moment is proportional to its *usefulness value* $pv(\cdot)$ at that moment, the bigger the former the smaller the latter, so

$$pc_a(x) = (1 - pv(x)) \cdot a \tag{9}$$

where $a \in [0, 1]$ is a normalizer constant.

4.3 Proposed personal utility (U)

Suppose the HO wants to do a good work, according to Equation (3), her personal utility $U(\cdot)$ will depend on whether a *positive action* (say *YES*) is taken or not. In this sense, one can define her utility $U(\cdot)$ as follows:

$$\begin{aligned} U(c|r) &= \psi(h, pc_a, p) \\ &= \begin{cases} \psi^+(h^+, pc_a, p), \text{for a positive action, and} \\ \psi^-(h^-, pc_a, p), \text{for a negative action} \end{cases} \end{aligned} \tag{10}$$

with,

$$\psi^+(h^+, pc_a, p) = (h^+ - pc_a) \cdot p + (h^+ - \lambda \cdot pc_a) \cdot w^+(p)$$

and,

$$\psi^-(h^-, pc_a, p) = (pc_a - h^-) \cdot (1 - p) + (pc_a - \lambda \cdot h^-) \cdot w^-(1 - p)$$

where, p is the current objective probability, $w(\cdot)$ is a *subjective probability weighting function*, and $\lambda > 1$ is the *loss-aversion coefficient*. Note that pc_a refers to the corresponding "positive framing" (cyan logistic curve in Fig. 4) of pv. The loss-aversion ("negative framing") is obtained with $\lambda \cdot pc_a$.

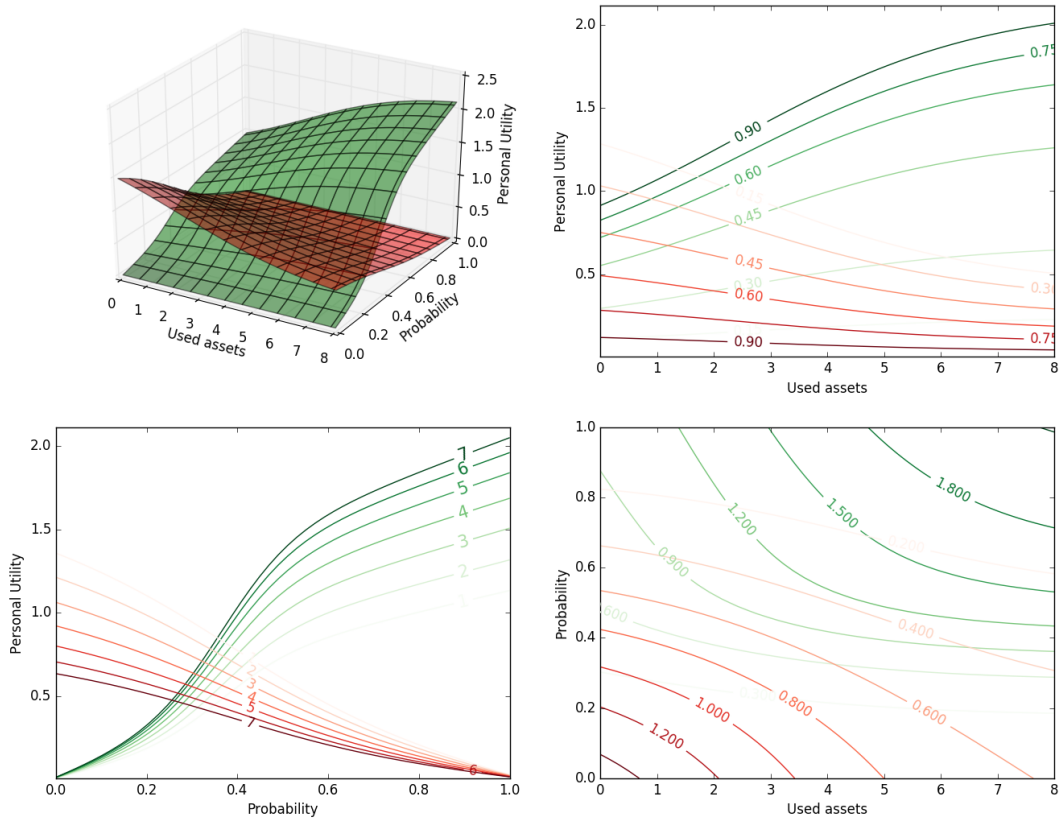

Figure 5: Personal utility in function of the framing for the earthquake scenario. The green color represents the positive framing ($\psi^+(\cdot)$) and the red color the negative framing ($\psi^-(\cdot)$).

The result generates two surfaces (see Figure 5 - top-left) for each scenario: one for the positive framing with the preference of saying "YES" (green - $\psi^+(\cdot)$) and other for the negative one with the preference of saying "NO" (red - $\psi^-(\cdot)$). The top-right plot in Figure 5 shows the personal utility as function of the used assets for a given probability value (color intensity). The bottom-left plot in Figure 5 shows the personal utility versus the probability for a given number of used assets. The bottom-right plot in Figure 5 represents the indifference curves (same utility lines) for both framings. These indifference curves are used to estimate the HO's decision, comparing the utilities of the framings (the greater one won).

4.4 Evaluation of the model

Suppose the HO wants to maximize her personal utility $U(\cdot)$, one can link an HO's decision to her $U(\cdot)$ by means of a maximization operator, and thus predicting HO's decision, a positive (pos) or a negative (neg) action, as follows:

$$\phi(h, \alpha, p) = \arg \max_{neg,pos} (\psi^-(h^-, pc_a, p), \psi^+(h^+, pc_a, p)) \tag{11}$$

Such a decision criterion should select which framing should be shown to HO in order to maximize the chance of having an appropriate decision, i.e., a decision aligned with the operational guidelines.

In order to validate the model obtained, a training subset with 75% of the data (homogeneously selected between the *text framings* and randomly among the individual data) is used for modeling and the remaining 25% is used to test the model. With the purpose of avoiding selection bias, a *repeated k-fold Cross Validation* is done (with $k = 4$ and 10 repetitions), in which the average values across all k trials is computed (Strimmer., 2015; Olsen, 2017).

The results comparing the decision criterion $\phi(.)$ over the test subset in order to predict HO's decision and the actual decisions made by the participants, are summarized in Table 3. In this Table the overall *Accuracy* rate is computed along with a 95% confidence interval, *Sensitivity*, also known as *Recall* is the number of positive predictions divided by the number of positive class values in the test data, and *PPV - Positive Predictive Value* is the number of positive predictions divided by the total number of positive class values predicted (Kuhn, 2008).

Table 3: Confusion Matrix and Statistics for HO's decisions prediction

Scenario	Probability range	Confusion matrix					
		Accuracy	Sensitivity	PPV	Prediction	Answer (%)	
						NO	YES
Earthquake	0 - 100%	0.8156	0.9264	0.8165	NO (Negative framing)	0.21	0.04
					YES (Positive framing)	0.14	0.61
	40 - 60%	0.7606	1.0	0.7605	NO	0.0009	0.00
					YES	0.23	0.76
Rock sampling	0 - 100%	0.8251	0.9374	0.7897	NO	0.31	0.04
					YES	0.14	0.51
	40 - 60%	0.7169	0.9698	0.6852	NO	0.15	0.02
					YES	0.26	0.57

These results suggest that the decision criterion $\phi(.)$ can predict HO's decision with a (considered) good accuracy. This means, $\phi(.)$ defines the framing that should be presented by the system, i.e., a positive framing for an expected "YES" answer else a negative one.

5 Discussion

In general, as pointed out in the introduction of this work, the Expected utility theory is used to model human decisions, as if the decision-maker was an unfailing machine with no time pressure to decide (Von Neumann and Morgenstern, 2007). However, Kahneman (2011) suggest that when DMs have to make crucial decisions under imperfect information conditions, humans are more prone to make foreseeable errors in judgment caused by *cognitive biases*. In particular, under situations where they are emotionally involved, as for instance, during space missions, emergencies (Robinette et al., 2016), natural disasters (Bevacqua et al., 2015), military operations (Schmitt et al., 2018) or any other complex and ambiguous environment (Barnes et al., 2015). In order to overcome these situations and to remain effective amid an unpredictable and diffused environment it is important to understand and deal with these *"hard-wired"* human processes (Klein, 1997). Thus, the present work proposed an interesting experiment and a formal approach to better understand human operator's decisions in operational contexts under uncertainty and emotional commitment.

According to the Prospect Theory (PT) authors (Kahneman and Tversky, 1979a), the rational decision to choose the best option among some risky or uncertain prospects doesn't depend on the maximum utility value, but on the human behavior, i. e., there is an asymmetry between gains and losses to be evaluated in case of dealing with risky situations. More precisely, the asymmetry in the evaluation of gains and losses lead participants to make different decisions for gains than for loss (Kahneman, 2011; Tversky and Kahneman, 1981; Kahneman and Tversky, 1979b). This asymmetry is also confirmed by the investigation of affective reactions associated with gains and losses. For instance, participants reported more positive affective reactions when they faced to gains compared to when they faced to losses (Stark et al., 2017). This asymmetry was also observed in the present study (see Table 1), which leads participants to favor gains when they faced to gains (positive framing) and to avoid losses when they faced to losses (negative framing), following then (Kahneman and Tversky, 1979b; Stark et al., 2017). Thus, it is worth to say that the Framing Effect (FE) is a strong and powerful cognitive bias in human decision making. Moreover, our results showed that the way the decision-making problem was presented to participants (positively or negatively framed) and the emotional commitment involved (saving lives vs. collecting rocks) statistically affected the choices made by the participants (see section 3.3). Table 1 shows that participants were more inclined to say "YES" in the Earthquake scenario than the Mars scenario. Moreover, the coefficients for *Probability* and used *Assets* in Table 1 also denote that increasing the probability value or the number of used assets also increase the willingness to say "YES". These factors also demonstrated to modulate the asymmetry between gains and losses (see figures 3 and 4) showing that such other factors would also influence the human operator's decisions in our experimental settings. The PT based utility function allowed to integrate such factors (see sections 4.1, 4.2.1 and 4.2.2)

Considering the prediction performance of the proposed model, some considerations could be discussed. Firstly, this *random intercept model* assumes that the fixed effect is the same for all subjects. However, this assumption cannot be totally validate, as different people respond differently in the same situation. Despite the fact that our study is limited to some extent to provide a generalization of the model, results are in line with previous studies showing that people differentially respond to FE. More precisely, individuals respond differentially in function of the emotions that they felt (Cassotti et al., 2012; Osmont et al., 2015) or if they have to decide for themselves or for others (Kappes et al., 2018). Thus, the *perceived value* could be more a *personal perceived value* than a common one. In line with this argument, it has

been shown that the perception of risk could be modulated by emotions (Lerner and Keltner, 2000). More precisely, angry participants perceived risk as more attractive than fearful participants. Thus, fearful participants are less loss aversive than angry participants (Lerner and Keltner, 2000). Secondly, some of the participants did not do the experiment in their first language, so, they might have misunderstood some information that was presented (as some of them reported at the end of the experiment), but it is important to notice that the GLMM analysis did not show any significant correlation with that variable. However, this could reduce the accuracy of the model. And, finally, others reported that, at the begin, they did not realize that there were different types of sentences and payed attention only in the probability value presented (*attentional tunneling*), leading to an incorrect situation awareness that likely resulted in taking a "wrong" decision.

6 Conclusion

As far as the authors know, this work is the first study where an HO's utility function based on the Prospect theory and a decisional model to predict those HO's decisions are proposed.The proposed model is based on the approach of multiple dimensions proposed by Kőszegi and Rabin (2006) in another context. But, contrary to their work, any simplification assumption (e.g. linear utility for gains and losses and no probability weighting function) is used in the present work. To consider a more general *gain-loss function*, the authors have considered a non linear probability weighting function $w(.)$ that respects the mathematical conditions of PT (as strictly increasing function), and have identified different coefficients from collected data to approach a personal perceived value function. The contributions of this work, should allow a cybernetic system to choose the framing to present to the human operator in order to induce the required *HO's decision* that is within the operational guidelines.

On one hand, different people respond differently in the same situation, while on the other hand same person responds differently in different situations. For the former case, and for increase prediction performance, the authors could explore tools from the *transfer learning literature* (Pan and Yang, 2010) to: (i) in a first step, cluster people who have a similar *personal perceived value*; (ii) applying the appropriate *perceived value* function, in accordance with this common behavior. For the latter case, the results of this study show that the two scenarios induced different behaviors among participants. It is reflected, in particular, on the resulting utilities functions. The perceived value and weighted probability functions were different across the two scenarios. Though one could say that the methodology is generalizable, each scenario model has to be fed with ad-hoc dataset. Consequently, more research must be conducted concerning new class of scenarios for potential overall model generalization.

Next step in this research, is programmed to evaluate the proposed model in a closed-loop operational situation, where the system should choose the framing to present to the HO, expecting to align her decision with the operational guidelines. The authors hope that this closed-loop system, which will integrate the human operator's utility function in a global criterion, maximizes the utility of the overall system compared to a system where no framed choice is automatically done.

References

Agresti, A. and Kateri, M. (2011). *Categorical data analysis*. Springer.

Bago, B. and De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition*, 158:90–109.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. *The Journal of Economic Perspectives*, 27(1):173–195.

Barnes, M. J., Chen, J. Y., and Jentsch, F. (2015). Designing for mixed-initiative interactions between human and autonomous systems in complex environments. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 1386–1390. IEEE.

Baron, J. (2007). *Thinking and Deciding*. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Belloni, A., Berger, A., Besson, V., Boissier, O., Bonnet, G., Bourgne, G., Chardel, P. A., Cotton, J.-P., Evreux, N., Ganascia, J.-G., et al. (2014). Towards a framework to deal with ethical conflicts in autonomous agents and multi-agent systems. In *CEPE 2014 Well-Being, Flourishing, and ICTs*, pages paper–8.

Bevacqua, G., Cacace, J., Finzi, A., and Lippiello, V. (2015). Mixed-initiative planning and execution for multiple drones in search and rescue missions. In *25th International Conference on Automated Planning and Scheduling (ICAPS)*, pages 315–323.

Biswas, M. and Murray, J. (2017). The effects of cognitive biases and imperfectness in long-term robot-human interactions: Case studies using five cognitive biases on three robots. *Cognitive Systems Research*, 43:266–290.

Cassotti, M., Habib, M., Poirel, N., Aïte, A., Houdé, O., and Moutier, S. (2012). Positive emotional context eliminates the framing effect in decision-making. *Emotion*, 12(5):926.

de Winter, J. C. and Dodou, D. (2014). Why the fitts list has persisted throughout the history of function allocation. *Cognition, Technology & Work*, 16(1):1–11.

Dehais, F., Causse, M., Vachon, F., and Tremblay, S. (2012). Cognitive conflict in human–automation interactions: a psychophysiological study. *Applied ergonomics*, 43(3):588–595.

Dehais, F., Goudou, A., Lesire, C., and Tessier, C. (2005). Towards an anticipatory agent to help pilots. In *AAAI 2005 Fall Symposium" From Reactive to Anticipatory Cognitive Embodied Systems", Arlington, Virginia*.

Dehais, F., Hodgetts, H. M., Causse, M., Behrend, J., Durantin, G., and Tremblay, S. (2019). Momentary lapse of control: A cognitive continuum approach to understanding and mitigating perseveration in human error. *Neuroscience & Biobehavioral Reviews*.

Dehais, F., Peysakhovich, V., Scannella, S., Fongue, J., and Gateau, T. (2015). Automation surprise in aviation: Real-time solutions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2525–2534. ACM.

Gombolay, M., Bair, A., Huang, C., and Shah, J. (2017). Computational design of mixed-initiative human–robot teaming that considers human factors: situational awareness, workload, and workflow preferences. *The International Journal of Robotics Research*, 36(5-7):597–617.

Guo, M., Andersson, S., and Dimarogonas, D. V. (2018). Human-in-the-loop mixed-initiative control under temporal tasks. *submitted to IEEE International Conference on Robotics and Automation, arXiv preprint arXiv:1802.06839*.

Jiang, S. and Arkin, R. C. (2015). Mixed-initiative human-robot interaction: Definition, taxonomy, and survey. In *Systems, Man, and Cybernetics (SMC), 2015 IEEE International Conference on*, pages 954–961. IEEE.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D. and Tversky, A. (1979a). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society*, pages 263–291.

Kahneman, D. and Tversky, A. (1979b). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91.

Kappes, A., Nussberger, A.-M., Faber, N. S., Kahane, G., Savulescu, J., and Crockett, M. J. (2018). Uncertainty about the impact of social decisions increases prosocial behaviour. *Nature human behaviour*, 2(8):573.

Klein, G. (1997). Developing expertise in decision making. *Thinking & Reasoning*, 3(4):337–352.

Kolling, A., Walker, P., Chakraborty, N., Sycara, K., and Lewis, M. (2016). Human interaction with robot swarms: A survey. *IEEE Transactions on Human-Machine Systems*, 46(1):9–26.

Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, pages 1133–1165.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Lerner, J. S. and Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & emotion*, 14(4):473–493.

Levin, I. P., Schneider, S. L., and Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, 76(2):149–188.

Matarić, M. J., Sukhatme, G. S., and Østergaard, E. H. (2003). Multi-robot task allocation in uncertain environments. *Autonomous Robots*, 14(2-3):255–263.

Murphy, R. R., Tadokoro, S., Nardi, D., Jacoff, A., Fiorini, P., Choset, H., and Erkmen, A. M. (2008). Search and rescue robotics. In *Springer Handbook of Robotics*, pages 1151–1173. Springer.

Nadendla, V. S. S., Brahma, S., and Varshney, P. K. (2016). Towards the design of prospect-theory based human decision rules for hypothesis testing. In *54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 766–773. IEEE.

Olsen, L. R. (2017). *groupdata2: Creating Groups from Data*. R package version 1.0.0.

Osmont, A., Cassotti, M., Agogué, M., Houdé, O., and Moutier, S. (2015). Does ambiguity aversion influence the framing effect during decision making? *Psychonomic bulletin & review*, 22(2):572–577.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Ren, P., Xu, Z., and Hao, Z. (2016). Hesitant fuzzy thermodynamic method for emergency decision making based on prospect theory. *IEEE Transactions on Cybernetics*.

Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 101–108. IEEE.

Schmitt, F., Roth, G., Barber, D., Chen, J., and Schulte, A. (2018). Experimental validation of pilot situation awareness enhancement through transparency design of a scalable mixed-initiative mission planner. In *International Conference on Intelligent Human Systems Integration*, pages 209–215. Springer.

Schurr, N., Marecki, J., and Tambe, M. (2009). Improving adjustable autonomy strategies for time-critical domains. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 353–360. International Foundation for Autonomous Agents and Multiagent Systems.

Souza, P. E., Chanel, C. P. C., Dehais, F., and Givigi, S. (2016). Towards human-robot interaction: a framing effect experiment. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 001929–001934. IEEE.

Stark, E., Baldwin, A. S., Hertel, A. W., and Rothman, A. J. (2017). Understanding the framing effect: do affective responses to decision options mediate the influence of frame on choice? *Journal of Risk Research*, 20(12):1585–1597.

Steiger, A. and Kühberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie*.

Strimmer., K. (2015). *crossval: Generic Functions for Cross Validation*. R package version 1.0.3.

Suarez, J. and Murphy, R. (2011). A survey of animal foraging for directed, persistent search by rescue robotics. In *Safety, Security, and Rescue Robotics (SSRR), 2011 IEEE International Symposium on*, pages 314–320. IEEE.

Suhonen, N. (2007). Normative and descriptive theories of decision making under risk: A short review. *Joensuu, Finland: University of Eastern Finland*.

Timotheou, S. and Loukas, G. (2009). Autonomous networked robots for the establishment of wireless communication in uncertain emergency response scenarios. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1171–1175. ACM.

Todd, P. M. and Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and brain sciences*, 23(05):727–741.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323.

Von Neumann, J. and Morgenstern, O. (2007). *Theory of games and economic behavior*. Princeton university press.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge university press.

Xue, S., Zeng, J., and Zhang, G. (2011). A review of autonomous robotic search. In *Electrical and Control Engineering (ICECE), 2011 International Conference on*, pages 3792–3795. IEEE.

Zhang, G. (2016). Comparison of decision-making mechanism between emotion behavior selection and prospect theory. In *8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, volume 1, pages 538–540. IEEE.