Preprints.org

Article

Not peer-reviewed version

Trust in Artificial Intelligence: Modelling Human Operators' Decision-Making in Highly Dangerous Situations

Alexander L. Venger and Victor M. Dozortsev *

Posted Date: 29 September 2023

doi: 10.20944/preprints202309.2099.v1

Keywords: human operator; trust in artificial intelligence; recommender systems; intelligent decision-making systems; admissible probability of disaster; and equipment predictive analytics

Preprints.org is a free multidiscipline platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Trust in Artificial Intelligence: Modelling Human Operators' Decision-Making in Highly Dangerous Situations

Alexander L. Venger [1] and Victor M. Dozortsev [2,*]

[1] Department of Social Sciences and Humanities, Dunba State University, Dubna, Moscow Region, 141982, Russia; venger.1@uni-dubna.ru
[2] Moscow Institute of Physics and Technology (MIPT), Moscow 117303, Russia
* Correspondence: dozortsev.vm@mipt.ru

Abstract: Here, we propose a prescriptive simulation model of a process operator's decision-making assisted by artificial intelligence (AI) algorithm in a technical system control loop. We analyze situations fraught with a catastrophic threat that may cause unacceptable damage. Operators' decision-making is interpreted in terms of a subjectively admissible probability of disaster and subjectively necessary reliability of its assessment. We distinguish four extreme decision-making strategies corresponding to different ratios between the above variables. An experiment simulating a process facility, an AI algorithm and operator's decision-making strategy was held. It showed that depending on the properties of a controlled process (the speed of hazard onset) and the AI algorithm characteristics (Type I and II error rate), each of such strategies or some intermediate strategy may prove to be more beneficial than others. The same approach is applicable to the identification and analysis of sustainability of strategies applied in real-life operating conditions, as well as to the development of a computer simulator to train operators to control hazardous technological processes using AI-generated advice.

Keywords: human operator; trust in artificial intelligence; recommender systems; intelligent decision-making systems; admissible probability of disaster; and equipment predictive analytics

1. Introduction

With the advent of modern IT tools, artificial intelligence has been steadily penetrating industrial automation. So far, it has been in the form of advice, which can be accepted or rejected by a human operator, and which relates to both avoiding undesirable operating modes of a facility as well as process equipment predictive analytics.

As an example, let us take a look at equipment predictive analytics when AI triggers an alert that requires shutting down a complex manufacturing process. Suppose the shutdown is very costly, but the potential accident AI warns about would lead to unacceptable damage (say, a nuclear power plant disaster). The operator can reject the AI advice and accept the corresponding risk or take the advice and shut down the process, but in the latter case if the alarm is false, there is a risk of significant losses due to the unnecessary shutdown.

Despite the operator's leading role in such a human-machine system (and possibly due to that role), the presence of AI gives rise to serious challenges related to workforce and production assets safety, staff motivation, ethics in industrial relations, etc. Along with the variety of factors (transparency of algorithms, responsibility for, and benefits of, the use of AI, etc.), operator's own decision-making strategy is determined by psychological factors, such as admissible probability of disaster and doubts about the accuracy of hazard assessments. Achieving a sufficient level of trust is a prerequisite for the survival and effective functioning of AI algorithms in a modern production environment. Thus, there is a growing urgency for psychological support in human-machine interaction involving AI.

© 2023 by the author(s). Distributed under a Creative Commons CC BY license.

The above problem is examined here in the context of human trust in technology against the background of enhancing intellectualization of technological systems. Our task was to test the hypothesis that there is a significant difference in the effectiveness of possible decision-making strategies for different facilities and AI algorithms.

2. The Problem of Human Trust in Complex Machinery

The trust of a human operator in AI cannot be dealt with in isolation from trust in machinery in general, an issue that emerged at the turn of the XVIII-XIX centuries at the time of nascent mechanization, when the Luddites were destroying power looms that were making British weavers' jobs redundant [1].

Proper recognition of cases when trust in, or on the contrary distrust of, technology devices is needed may have a fateful significance. A tragic precedent is the 1941 Japanese air attack on Pearl Harbor, a United States naval base, when fuzzy signals from the then imperfect radars were not perceived as a genuine threat, and an ill-conceived decision-making protocol thwarted the response. In contrast, a prudent distrust of technology saved the USSR and the United States from a nuclear clash at the height of the Cold War. Indicatively, in the early 1980s, more than 10 false positive alarms per day were recorded by the American side alone. All of them were the result of malfunctions, hardware and software failures, or natural interference [2].

With the advance of technology, the problem of trust became increasingly multidimensional and philosophers, literary figures, sociologists, culturologists, psychologists, and science fiction writers were getting involved in it. The adoption of computer-integrated manufacturing affected a new important category of users – industrial system operators.

The problem of trust in technology is a traditional topic for psychological research [3]. The fundamental provision is that the level of trust should match technological capabilities. A mismatch may lead to overtrust or over-distrust posing a threat of either a decrease in safety or an unjustified rejection of the benefits of modern automation. At the same time, in addition to the technical parameters of automation itself (reliability, safety, ease of use, etc.), the level of trust is influenced by the human operator profile: experience, professional competence, self-esteem, and other personal traits [4].

Nowadays, human operators must increasingly embrace recommender systems and intelligent decision support that have become indispensable for high-quality control of complex technological facilities. Notable practical applications of AI systems include laser sintering of metal products, proactive recognition of conveyor belt wear and tear or burnout of foundry ladles lining, and predictive equipment maintenance, which radically mitigates the risk of damage and the threat to health and life at work. There is no alternative to AI methods in terms of equipment maintenance in case of exceptionally high breakdown costs (*e.g.*, steel pipe welding equipment or blast furnace compressors). While in some cases, an operator has enough time (hours or days) to analyze AI advice, in other cases, a signal may arrive immediately before a possible failure, giving an opportunity to save critical equipment but also requiring a prompt and responsible decision from the human operator.

3. Specifics of Trust in/Distrust of IA Systems

Motivational trap. If correct AI advice is rejected or the wrong advice is accepted, the human operator risks revealing a skills gap. If correct advice is accepted, then AI will prove to be at least faster, or even "smarter" than the human operator. Only rejecting the wrong advice would benefit a human operator. Given operator's formal and often informal responsibility and status as a hired employee, such "asymmetrical" motivation does not boost trust towards AI.

Reluctance to use AI. The problem of reluctance to use technological innovations is also kind of eternal. An obvious tool to improve staff willingness is training and re-profiling of users and professionals, among other things, with reliance on a technological facility high-precision simulation. However, in case of AI, the situation is exacerbated by a number of additional factors:

- AI tools are still quite "young", not insured against "teething problems" and can behave unpredictably in case of situations that have not been covered during the algorithms learning stage;
- Staff fears of losing jobs were heightened at every critical point in the development of automation: during transition from analog to digital computerized control systems, introduction of the first model-based predictive control systems ("industrial autopilots") and, finally, upon penetration of AI tools into automation;
- operators' AI-related concerns are often complemented by their reluctance to share one of the principal human advantages—the ability to think—with a machine [5]. Now, automation claims not only fast routine tasks that are beyond human reach but also optimization, planning, predictive analytics, i.e., it increasingly encroaches on the "sancta sanctorum", which is widely viewed as accessible to natural human intelligence only.

4. Simulation Model of an AI-Based Human-Machine System

We have already proposed a generalized description of decision-making in situations characterized by the risk of catastrophic consequences [6]. Psychological research predominantly examines risk, which comes down to undesirable but not radical losses: "Risk is the potential that a decision will lead to a loss or an undesirable outcome" [7] (p. 3). In contrast, we looked at the risk of an event leading to unacceptable damage. As is customary in operator accident prevention training, the threats of such events are repeatedly simulated, for example, on computer simulators.

We investigated the influence of trait anxiety, which according to our hypothesis, is represented by two independent parameters, conventionally known as "apprehension" and "doubt". Apprehension characterizes the probability of a catastrophe that a person considers admissible for himself. "Doubt" is a subjectively assessed degree of reliability of one's own assessment of such probability, necessary for making a final decision.

When building the model, we relied on the sequential decision theory [8] limiting ourselves to choosing between two possible final decisions and an interim one:

- D^{frw}—taking a risky action, continuing the process despite the existing probability of an accident (move forward);
- D^{stp}—refusal to take a risky action, stopping the process, which has a fairly high price;
- D^{tst}—a significantly less costly interim decision: gathering additional information as a basis for the final decision D^{frw} or D^{stp}.

It was assumed that the choice of one or another option is determined by the assessment of probability (subjective probability) of catastrophic consequences in relation to the specified parameters of trait anxiety of the decision-maker. With growing "apprehension", there is an increasing tendency to choose a D^{stp} decision (refusal to take a risky action). A high level of "apprehension" helps avoid a disaster but reduces benefits ("gains") that a risky but successful action could generate.

With increasing "doubt", there is a growing tendency to choose an interim D^{tst} decision (to gather additional information). It also helps avoid a disaster but given the high cost of information-gathering it significantly reduces the total gain. In addition, a D^{tst} decision becomes meaningless, if there are no sources of additional information, there is no time to gather it, or the information gathering process itself is fraught with a considerable risk.

In the proposed model, an operator observes a dynamic process. The process periodically falls into a state that may potentially lead to a catastrophic event of which AI gives an early warning to the operator. AI recommendations are not perfect: there are Type I ("false alarms") as well as Type II ("missed targets") errors. The operator can stop the process (D^{stp}) or keep it running (D^{frw}) and can also make an interim decision to obtain additional information (D^{tst}).

The model is based on singling out three zones:

1. Risk acceptance zone: according to subjective assessment, the risk is not higher than admissible, and the subjective reliability of the assessment is sufficient. A risky D^{frw} decision to keep the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

process running is made. (If AI recommended stopping the process, then such recommendation is rejected.)

2. Zone of uncertainty: according to subjective assessment, the risk is not higher than admissible, but the subjective reliability of the assessment is insufficient. An interim D^{tst} decision to collect additional information is made. If, as a result, sufficient reliability of the initial assessment is achieved, then a D^{frw} decision to keep the process running is made. Otherwise, a D^{stp} decision to stop the process is made. (AI recommendation is accepted.)
3. Excessive risk zone: according to subjective assessment, the risk is higher than admissible. A D^{stp} decision to stop the process is made. (AI recommendation is accepted or a proactive decision is made based on operator's own assessment of the state of the process).

The model includes the following components:

I. "Process";

II. "AI" observing the process and predicting its state at the next two timepoints; and

III. "Operator" who has an opportunity to make one of the three decisions – D^{frw}, D^{stp}, or D^{tst} – at any given time.

Below is a detailed description of each block.

I. The following time-discrete dynamic process $\{X_i\}$ is modelled:

$$X_i = \max\{0;\ a^*X_{i\text{-}1} + b^*z_i\};$$

$$X_0 = ½\Delta,$$

where

Δ is the value, which if exceeded, is treated as a "catastrophe" resulting in unacceptable damage;

$z_i \sim N\ (0, 1)$ is a standard normally distributed random variable;

$a\in(0, 1]$, $b > 0$ are constants that determine the dynamics and the power of the process (and as a result, the frequency and suddenness of the onset of a catastrophic risk);

$i\in[1, n]$; n—process duration (the total number of steps).

II. According to the model, in addition to the value of X_i, AI also "knows" the predicted values of X'_{i+1} and X'_{i+2}; if $X'_{i+1} > \Delta$ or $X'_{i+2} > \Delta$, AI triggers an alarm (suggests that the process be stopped). "False alarms" (FA) and "missed targets" (MT) are also possible. The probabilities of each of these scenarios are determined by natural numbers predetermined by the researcher $0 \leq M^{FA} < M^{MT} \leq 1000$, in correlation with the values of the random variable g_i evenly distributed over the segment [1; 1000]:

- $g_i \leq M^{FA}$: the signal is given regardless of the values of X'_{i+1} and X'_{i+2}, which generates "false alarms", but sometimes it can accidentally coincide with a correct warning;
- $M^{FA} < g_i \leq M^{MT}$: the signal is given, if $X'_{i+1} > \Delta$ or $X'_{i+2} > \Delta$;
- $g_i > M^{MT}$: there is no signal, regardless of the values of X'_{i+1} and X'_{i+2}, which can generate a "missed target".

Thus, at $M^{FA} = 0$ there are no "false alarms" and at $M^{MT} = 1000$ there are no "missed targets" (**Figure 1**). If a signal is given at the i-th step, then at the next (i+1) step there is no signal.

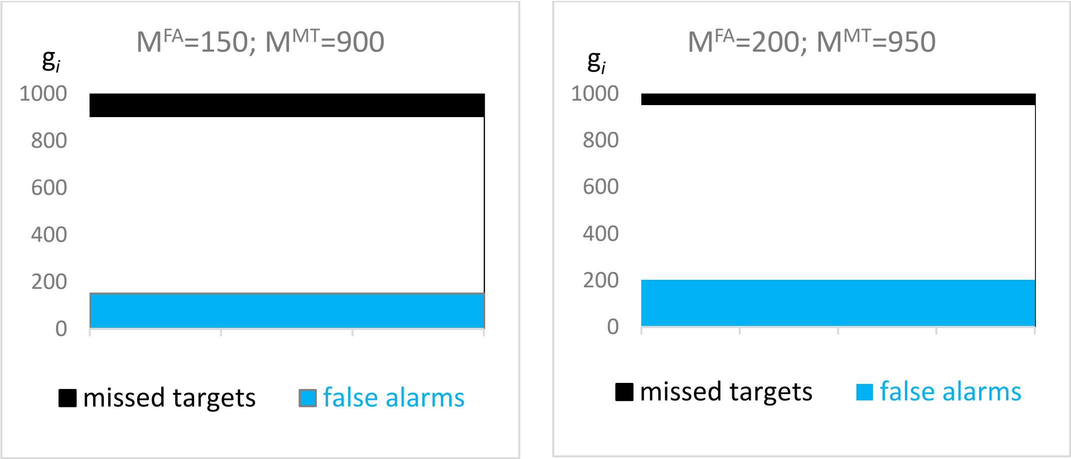

Figure 1. Zones of false alarms and missed targets at different values of M^{FA} and M^{MT}.

III. The model assumes that the operator is guided by both AI signals and his own assessment of the state of the process. He, however, is not able to determine the exact value of X_i but only the boundaries of the range [$Y^{btm}{}_i$, $Y^{top}{}_i$] in which it is located. The boundaries are set as follows:

$$Y^{btm}{}_i = X_i + k^*(wi - 2.5);$$

$$Y^{top}{}_i = X_i + k^*(w_i + 2.5),$$

where

$k > 0$ is a constant and w_i is a random variable with truncated standard normal distribution with truncation levels at [–2.5, +2.5].

The value of $Y_i = ½\ (Y^{btm}{}_i + Y^{top}{}_i)$ is a subjective assessment of the state of the X_i process; and the constant k determines the degree of accuracy of such assessment.

It is assumed that when *an AI alarm signal arrives*, the operator follows the algorithm using the constants $0 < δ^{frw} \le δ^{stp}$, which characterize the individual strategy (the higher is the level of "apprehension", the lower is $δ^{stp}$; the higher is the level of "doubt", the lower is $δ^{frw}$):

- if $Y_i \le δ^{frw}$, then a D^{frw} decision is made, *i.e.*, the [0, $δ^{frw}$] segment is the *risk acceptance zone*;
- if $δ^{frw} < Y_i \le δ^{stp}$, then a D^{tst} decision is made, *i.e.*, the ($δ^{frw}$, $δ^{stp}$] interval is the *zone of uncertainty* (in case of $δ^{stp} = δ^{frw}$ it is absent);
- if $Y_i > δ^{stp}$, then a D^{stp} decision is made, *i.e.*, any value higher than $δ^{stp}$ is the *excessive risk zone*.

A D^{frw} decision is based on the belief that the AI signal was a "false alarm". If the signal was indeed false, then the process moves one step forward. If the signal was correct, *i.e.*, $X'_{i+1} > Δ$ or $X'_{i+2} > Δ$, then a "catastrophe" occurs.

A D^{stp} decision is based on the notion that the probability of a potential catastrophe is excessively high. It means stopping the process and then bringing it back to the initial level: $X_{i+1} = 0$; $X_{i+2} = ½Δ$.

A D^{tst} interim decision means taking an "exploratory step": $X_i \Rightarrow X_{i+1} = X'_{i+1}$. The final decision is determined by the value of Y_{i+1} (see **Figure 2**):

- If $Y_{i+1} \le δ^{frw}$, then a D^{frw} decision is made; $X_{i+1} \Rightarrow X_{i+2} = X'_{i+2}$;
- If $Y_{i+1} > δ^{frw}$, then a D^{stp} decision is made; $X_{i+2} = 0$; $X_{i+3} = ½Δ$.

Two "exploratory steps" in a row are impossible.

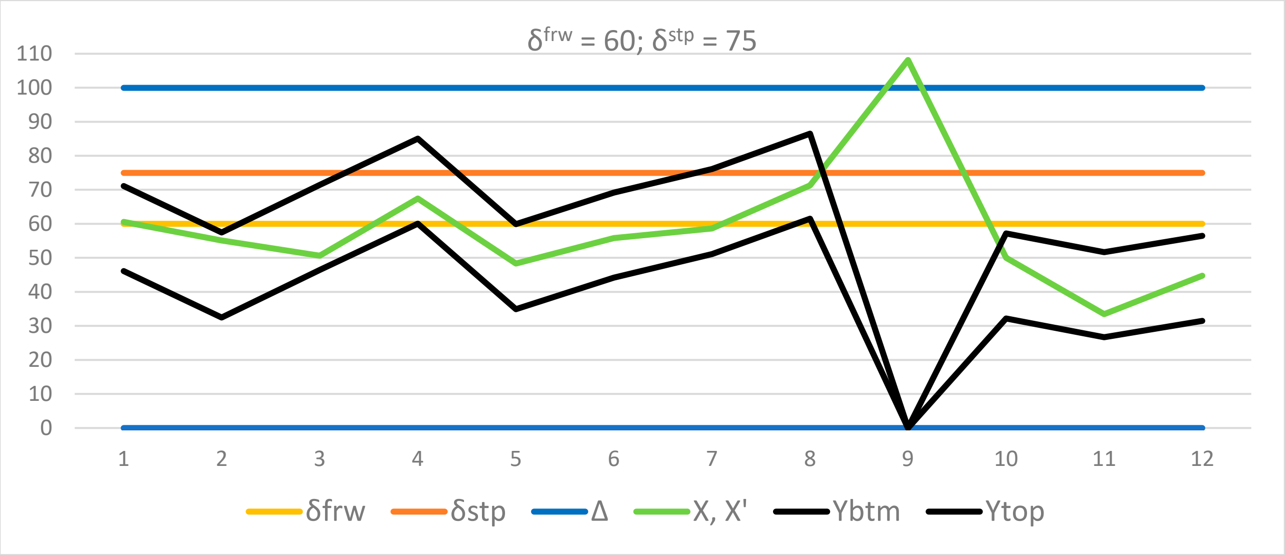

Figure 2. Alarm signal at step #7; $Y^{btm}{}_7 = 51.1$; $Y^{top}{}_7 = 76.1$; $Y_7 = ½(Y^{btm}{}_7 + Y^{top}{}_7) = 63.6$; $δ^{frw} < Y_7 < δ^{stp} \Rightarrow D^{tst}$ decision; $Y^{btm}{}_8 = 61.5$; $Y^{top}{}_8 = 86.5$; $Y_8 = ½(Y^{btm}{}_8 + Y^{top}{}_8) = 74.0$; $Y_8 > δ^{frw} \Rightarrow D^{stp}$ final decision.

A possible way to set such differences between D^{tst} and D^{frw} decisions on a simulator is to slow down the process when a D^{tst} decision is made to allow the operator to stop the process at the next step if necessary; if AI advice is rejected (a D^{frw} decision), the speed of the process is too fast to stop it at the next step.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

In the absence of an alarm, the operator is guided only by his own assessment of the state of the process. The algorithm of his actions is similar to the previous one, but instead of the constants δ^{frw}, the constants δ^{stp} are used:

$$\delta_h^{frw} = \delta^{frw} + h^*(\Delta - \delta^{frw});$$

$$\delta_h^{stp} = \delta^{stp} + h^*(\Delta - \delta^{stp}),$$

where

$h \in [0, 1]$ is a parameter that reflects the level of trust of the AI operator (the degree of his confidence that there are no "missed targets").

If $h = 0$, then $\delta_h^{frw} = \delta^{frw}$; $\delta_h^{stp} = \delta^{stp}$.

If $h = 1$, then $\delta_h^{frw} = \delta_h^{stp} = \Delta$.

With full confidence that AI has no "missed targets" ($h = 1$), the operator does not stop the process or take "exploratory steps" without an AI signal. With complete distrust of AI ($h = 0$), operator's actions are the same whether there is a signal or not. At intermediate values of h, the *zone of risk acceptance* is wider, and the *zone of excessive risk* is narrower compared to the respective zones in case of an AI alarm. (**Figure 3**). In other words, subject to operator's trust in AI reliability in determining the target, the absence of a signal is used by the operator as additional information that enables him to make bolder decisions.

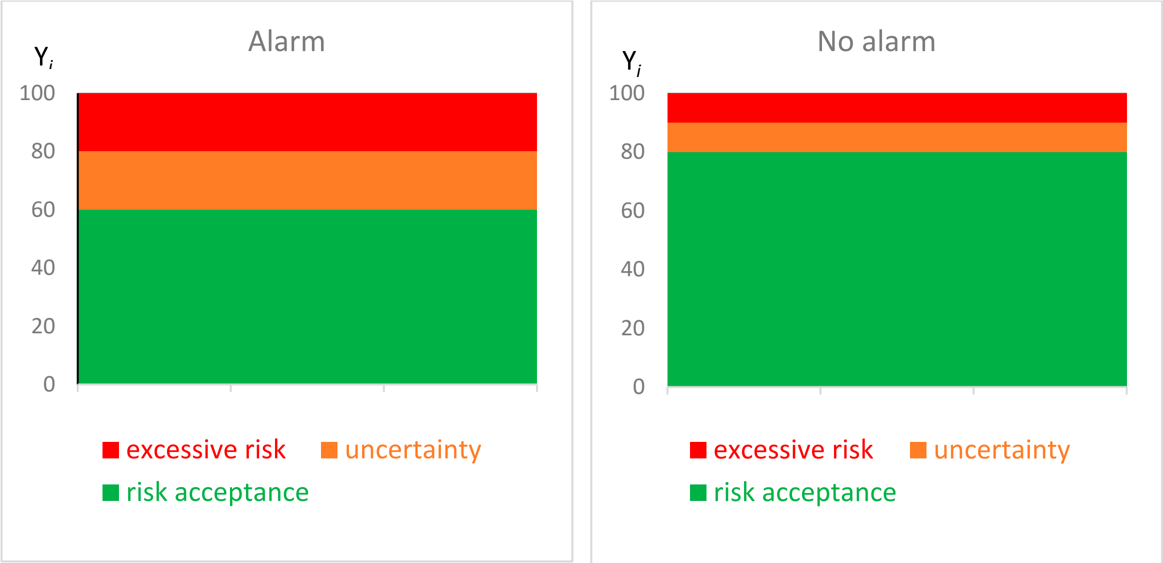

Figure 3. Zones of risk acceptance, uncertainty, and excessive risk depending on the presence or absence of an alarm signal ($\delta^{frw} = 60$, $\delta^{stp} = 80$; $h = 0.5$; $\delta_h^{frw} = 80$, $\delta_h^{stp} = 90$).

The model envisages a fine charged for stopping the process u^{stp} and for performing a verification u^{tst}. There is a reward prz^{FA} for false alarm identification (continuing the process, despite an erroneous AI signal), and a reward prz^{MT} for proactive shutdown of the process in the event of an actual threat of an accident in the absence of an AI signal. In addition to minimization of the probability of an accident, there is an *integral indicator* U, which serves as a criterion of operator's actions success and is the sum total of fines and bonuses accrued by the end of the process (n-th step):

$$U = \sum(prz^{FA} + prz^{MT} - u^{stp} - u^{tst}),$$

which, generally speaking, can be negative.

As appears from the above description, simultaneously high values of δ^{stp} and δ^{frw} are characteristic of an *extreme risk* strategy. High values of δ^{stp} with moderate values of δ^{frw} represent *a moderate risk strategy.* Moderate values of δ^{stp} and low values of δ^{frw} evidence a *moderate risk avoidance* strategy. Simultaneously low values of δ^{stp} and δ^{frw} are *an extreme risk avoidance* strategy.

In operators' daily activities, one may expect intermediate, rather than extreme, strategies more or less similar to the above options.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

5. Outcomes

The model has been tested to compare the predicted effectiveness of different operator strategies. Therefore, parameters δ^{frw} and δ^{stp} characterizing such strategies were varied. The remaining parameters were represented by two contrasting values (with a low and high objective probability of process emergency states) or remained unchanged. So, during the testing process, the following parameters did not change:

1. $n = 1000$ (*i.e.*, the overall duration of the process was 1000 steps);
2. $\Delta = 100$ (respectively, the initial value of $X_0 = 50$; it was reset each time after the process was shut down);
3. $a = 0.99$ (high process inertia);
4. $M^{FA} = 300$, *i.e.*, in 30% of cases, an AI signal was given regardless of either the current or the predicted state of the process[1];
5. $u^{stp} = 200$ (process shutdown cost);
6. $prz^{FA} = 100$ (reward for "false alarm" detection by the operator);
7. $prz^{MT} = 300$ (reward for proactive shutdown of the process if there is a treat of an accident without an AI signal).

The average frequency of accidents *Crsh* and the integral indicator *U* were checked. Two scenarios were simulated:

1. Allowing, as we assumed, a risk strategy: with rare real threats of an accident ($b = 7$; the average number of threats per 1000 steps M = 1.41 while the frequency of "false alarms" was much higher), with a high accuracy of subjective assessment of the state of the process ($k = 1$), without AI Type II errors ("missed target") ($M^{MT} = 1000$) and with operator's confidence in the absence of such errors ($h = 1$). This option envisioned a low cost of verification ($u^{tst} = 1$).
2. Requiring much more caution: with more frequent accident threats ($b = 10$; average number of threats per 1000 steps M = 2.64), reduced accuracy of subjective assessment of the state of the process ($k = 5$) and the possibility of AI Type II errors ($M^{MT} = 800$). Cases with a low verification cost ($\mathrm{u}^{tst} = 1$; the integral indicator was designated as U_1) and a high cost ($u_{tst} = 50$; the integral indicator – U_2) were tested.

With each combination of parameters, we ran a series of 200 samples of 1,000 steps each. The values of the individual parameters δ^{frw} and δ^{stp} changed in 5-unit increments. The minimum reasonable value of δ^{frw} was $\delta^{frw} = 55$ since the initial state of the process was $X_0 = 50$. The values of $\delta^{frw} = 55$ and $\delta^{frw} = 60$ were considered low (*risk avoidance* strategies), $\delta^{frw} = 65$ and $\delta^{frw} = 70$ were considered medium, and $\delta^{frw} = 75$ and above were considered high (*risk strategies*).

$Q = 0.05$ was assumed to be the maximum permissible frequency (probability) of an accident. Under normal circumstances, such high probability of a catastrophic accident would be unacceptable. For example, in construction in most countries, the maximum allowable annual individual risk associated with natural disasters is from 10^{-2} to 10^{-3} [9]. When a group (social) risk rather than an individual risk is considered, its permissible probability is sharply reduced. However, in a pre-emergency situation simulated by us, which is repeatedly run on simulators we considered the probability of $q = 0.05$ to be acceptable.

In the first ("encouraging" risk) scenario, the optimal value of δ^{frw} proved to be $\delta^{frw} = 75$. At higher values of δ^{frw}, the average frequency of "accidents" exceeded the maximum permissible level. So, already at $\delta^{frw} = \delta^{stp} = 80$, it was 0.080 per 1000 steps. At $\delta^{frw} < 75$, the integral indicator *U* significantly decreased. Thus, at $\delta^{frw} = 70$, the maximum result reaches 19,759 (at $\delta^{stp} = 85$), which is significantly lower than the result in the pair ($\delta^{frw} = 75$, $\delta^{stp} = 85$), which is 20,242 ($p < 0.001$; hereinafter, the significance of the differences was checked by Student's *t*-test).

The average frequency of accidents Crsh and the integral indicator *U* (M is the mean value, S is standard deviation) at $\delta^{frw} = 75$ and variable values of δ^{stp} are set out in **Table 1**.

[1] High frequency of "false alarms" is important when training operators to recognize them and is, therefore, typical in process simulations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

Table 1. Accident Frequency and integral indicator at $b = 7$, $k = 1$, $\delta^{frw} = 75$.

		δ^{stp}				
		75	**80**	**85**	**90**	**95**
Crsh		0.015	0.030	0.050	0.045	0.050
U	M	19,903	20,036	**20,242**	20,159	20,132
	S	1,035	1,027	1,002	975	1,037

As it appears from the Table, the optimal combination was ($\delta^{frw} = 75$, $\delta^{stp} = 85$). The difference between the integral indicator *U* in pairs ($\delta^{frw} = 75$, $\delta^{stp} = 85$) and ($\delta^{frw} = 75$, $\delta^{stp} = 75$) is statistically significant at $p < 0.001$.

In other words, under given conditions, the optimal strategy is intermediate between the strategies of *extreme risk* and *moderate risk*: a combination of a fairly wide zone of risk acceptance (a low level of "doubt") with an average width of the zone of uncertainty (an average level of "apprehension") and, accordingly, a narrow zone of excessive risk. Further narrowing of the zone of excessive risk did not lead to an increase in the integral indicator (**Figure 4**).

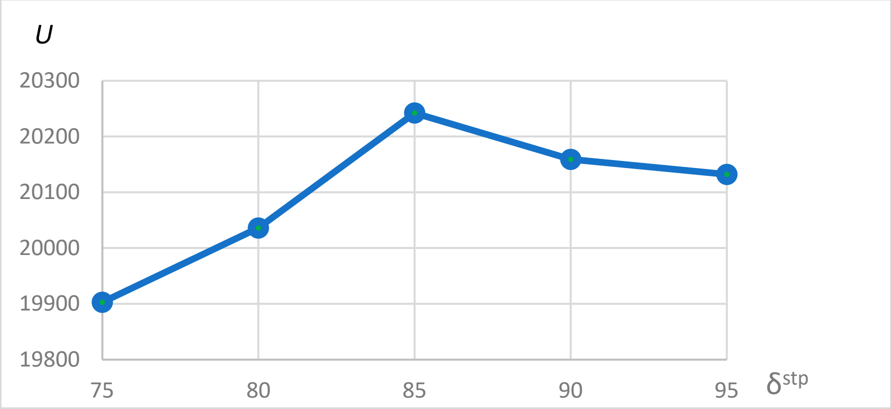

Figure 4. Integral indicator *U* at $\delta^{frw} = 75$ depending on δ^{stp}.

With more stringent requirements to the permissible probability of an accident, a more cautious strategy should be chosen, *i.e.*, the value of δ^{frw} is decreased (the risk acceptance zone is narrowed). Thus, at $\delta^{frw} = \delta^{stp} = 70$, the frequency of accidents per 1000 steps with 400 samples was $q = 0.005$. When checking the pair ($\delta^{frw} = 65$, $\delta^{stp} = 90$), there was not a single accident in a series of 400 samples of 1000 steps each. It is a strategy, intermediate between the strategies of *moderate risk* and *moderate risk avoidance*.

In the second scenario of the parameter values (with a significantly higher risk of process emergencies and, in particular "missed targets "), high confidence in AI (h =1) did not result in the required minimum of accidents even at minimum values of δ^{frw} and δ^{stp}. So, at $\delta^{frw} = \delta^{stp} = 55$ the average number of accidents per 1000 steps was 0.11. Successful completion of the process by the operator was observed at a significantly lower confidence level of $h = 0.40$. The maximum allowable pair values of δ^{frw} and δ^{stp} in this case were ($\delta^{frw} = 60$, $\delta^{stp} = 70$) and $\delta^{frw} = \delta^{stp} = 65$. At higher values, there was an excessive frequency of accidents of 0.065 at ($\delta^{frw} = 55$, $\delta^{stp} = 75$), and of 0.105 at ($\delta^{frw} = 65$, $\delta^{stp} = 70$).

The frequency of accidents and the integral indicator for different permissible combinations of parameters are set out in **Table 2**.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

Table 2. Accident frequency and integral indicator at $b = 10$, $k = 5$, $h = 0.40$.

δ^{frw}		55				60			65
δ^{stp}		55	60	65	70	60	65	70	65
Crsh		0.000	0.010	0.010	0.015	0.015	0.020	0.030	0.020
U_1	M	11,293	13,333	14,285	14,628	13,914	15,391	**15,955**	15,678
	S	2,436	2,082	2,202	1,893	2,182	1,693	1,648	1,880
U_2	M	11,293	12,627	13,152	13,151	13,914	14,867	15,017	**15,678**
	S	2,436	2,186	2,452	2,169	2,182	1,781	1,832	1,880

As it appears from the Table, with a low cost of verification, *a moderate risk avoidance strategy* is optimal: $\delta^{frw} = 60$, $\delta^{stp} = 70$ (**Figure 5**).

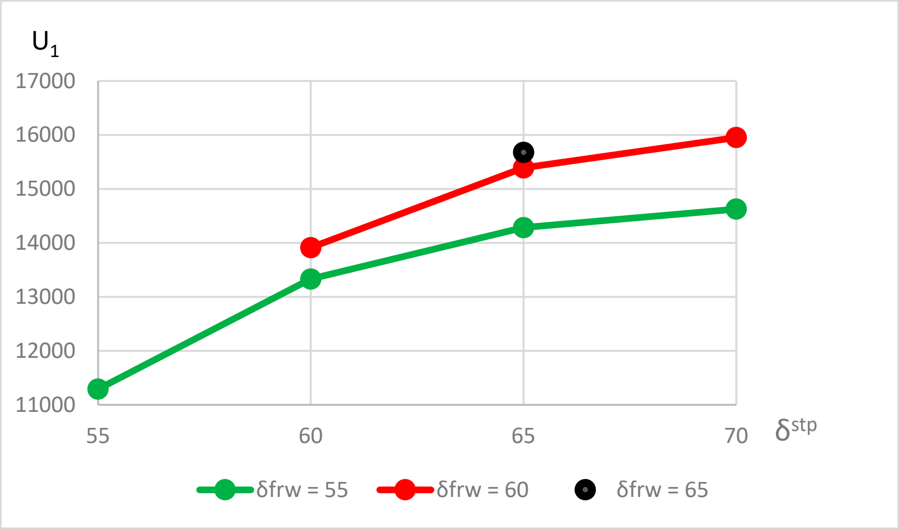

Figure 5. Integral indicator U_1 at different values of δ^{frw} and δ^{stp}.

In this scenario, the integral indicator statistically significantly (at $p < 0.001$) exceeds the cumulative sums in each of the other scenarios with accident frequency not exceeding 0.05. It is a scenario with a narrow risk acceptance zone and an average width of the uncertainty zone (an average level of “apprehension” and a high level of “doubt”).

With growing verification cost, the optimal frequency of verification predictably decreases, *i.e.*, as here, the zone of uncertainty becomes narrower or disappears altogether. The optimal solution is ($\delta^{frw} = \delta^{stp} = 65$), which provides a significantly higher U_2 than each of the other reviewed scenarios ($p < 0.001$) (**Figure 6**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 September 2023 doi:10.20944/preprints202309.2099.v1

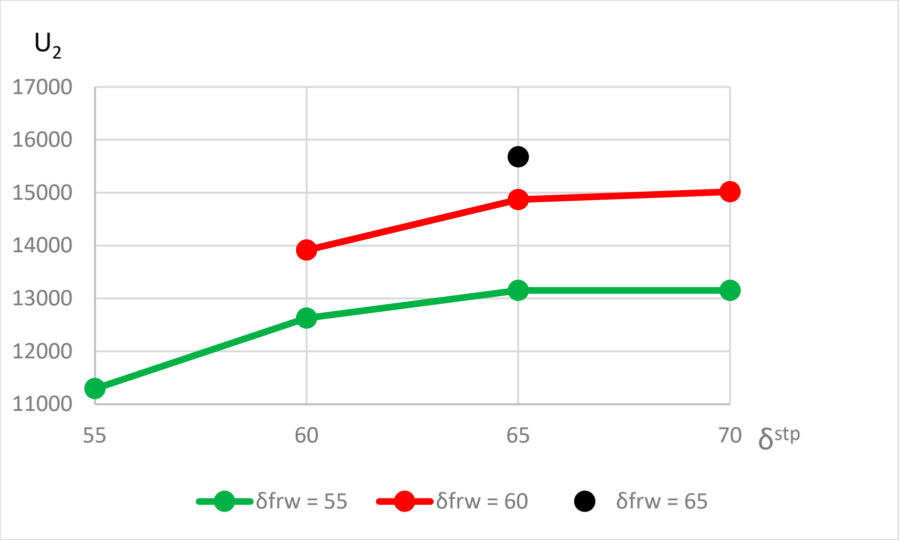

Figure 6. Integral indicator U_2 at different values of δ^{frw} and δ^{stp}.

6. Discussion

As expected, the model predicts a pronounced dependence of the effectiveness of a particular operator's strategy on the process dynamics (frequency of accident threats) and the characteristics of AI signals (frequency of "missed targets"). The optimal strategies under certain conditions were relatively pure strategies of *moderate risk avoidance* (with a high level of "doubt" and an average or low level of "apprehension") and *extreme risk avoidance* (with high levels of both parameters).

The strategy *of moderate risk avoidance* (low δ^{frw} and medium δ^{stp}) proved to be optimal with a relatively high frequency of threats of accidents, the presence of AI "missed targets", inaccurate subjective assessment of the process state by the operator, a not-too-high cost of gathering additional information, and pretty low safety level requirements (when maximum permissible probability of an accident was q = 0.05 per 1000 steps). The strategy of *extreme risk avoidance* (low values of both δ^{FRW} and δ^{STP}) becomes optimal with higher safety requirements, *i.e.*, when the value q goes down (*e.g.*, to q = 0.01).

The strategy *of extreme risk* with low levels of both "apprehension" and "doubt" (high values of δ^{frw} and δ^{stp}) in its pure form did not turn out to be optimal for any of the tested parameters. An intermediate strategy between *extreme risk* and *moderate risk* strategies with a low level of "apprehension" (high values of δ^{stp}) and a slightly higher level of "doubt" (lower value of δ^{frw}) proved to be optimal in case of a rare threat of accidents, absence of AI "missed targets", high accuracy of subjective assessment of the state of the process by the operator and low requirements to the safety level (q = 0.05). As safety requirements increase, a more cautious strategy, intermediate between *moderate risk* and *moderate risk avoidance strategies* becomes optimal.

Thus, the model predicts that the optimal level of operator's trait anxiety with respect to each of the two parameters identified by us is different depending on the dynamics and capacity of the simulated process and existence / absence of AI "missed targets". At the same time, some studies have shown that there is an individual level of anxiety that ensures top performance by an individual, in particular highest achievements in sports [10]. It can be assumed that the highest efficiency would be achieved by an operator when this level is adequate to the process dynamics and the specific features of AI recommendations.

7. Conclusions

There is no alternative to artificial intelligence in an increasingly wide range of tasks, including industrial automation. AI tools affecting the safety of people, production assets and infrastructure are proactively introduced. Operators' reluctance to use AI is largely driven by the insufficient level of trust in AI, which cannot be improved unless subjective factors and individual psychological profiles are considered.

Our simulation experiment validated the hypothesis that the degree of operator's success may depend on various combinations of parameters of admissible probability of disaster and the subjectively necessary reliability of its assessment. These findings unlock opportunities for future research going beyond the mathematical modeling of decision-making per se. Thus, the proposed model can be used in a psychological experiment to determine the propensity of operators to a particular strategy. It would make it possible to trace possible changes in the strategy of operators based on their work with the simulation model under different human-machine system conditions.

The findings also open up prospects for the development and reinforcement of operator's AI system skills through training and re-training on the basis of proven computer simulators, including a high-fidelity model of technological system (an actual process facility and a control system) and an advice-generating AI algorithm. Based on a field-proven decision-making simulation model, such training could consider individual operators' personality traits to identify preferred strategies, the level of required information support (awareness of AI advice generation mechanisms and accuracy boundaries, and the consequences of their acceptance or rejection), the format of offered advice, the level of detail, explanation, and justification. All the above would enhance operators' trust in AI systems, ensure their mutual adaptation and harmonization of human-machine interaction.

References

1. Jones S.E. Against technology: from the Luddites to Neo-Luddism. N.Y.: Taylor & Francis, 2006. [Google Scholar G]
2. Hart, G., Goldwater, B. Recent False Alerts from the Nation's Missile Attack Warning System. Washington: U.S. Government Printing Office, 1980.
3. Lee J., See K. Trust in technology: Designing for appropriate reliance. Human Factors. 2004. V. 46. No. 1; pp. 50–80.
4. Akimova A., Oboznov A. The factors of increase in trust and decrease in distrust of human to technique. Psychological Studies. 2017. 10(53). DOI: https://doi.org/10.54359/ps.v10i53.369.
5. Alekseev, A., Garbuk, S. How can you trust Artificial Intelligence Systems? Objective, Subjective and Intersubjective parameters of Trust. Artificial societies. 2022. V. 17, No. 2. DOI: https://doi.org/10.18254/S207751800020550-4
6. Venger, A.L. Mathematical model of decision making in extreme situations. Automation in industry. 2018. No. 6; pp. 32–36.
7. Lu, J., Jain, L. C., & Zhang, G. Risk management in decision making. Handbook on Decision Making: V. 2: Risk Management in Decision Making. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012; pp. 3–[https://scholar.google.com/scholar?hl=ru&as_sdt=0%2C5&q=Lu%2C%C2%A0J.%2 G].
8. LaValle, S.M. Sequential Decision Theory. Planning Algorithms. Cambridge University Press, 2006. Chapt. 10, pp. 495-559.
9. Sim, K.B., Lee, M.L. & Wong, S.Y. A review of landslide acceptable risk and tolerable risk. Geoenviron Disasters, 2022. **9**, 3. https://doi.org/10.1186/s40677-022-00205-6.
10. Ruiz, M. C., Raglin, J. S., & Hanin, Y. L. The individual zones of optimal functioning (IZOF) model (1978–2014): Historical overview of its development and use. International Journal of Sport and Exercise Psychology, 2017, 15(1), 41–63. https://doi.org/10.1080/1612197X.2015.1041545.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.