Article

Trust in Artificial Intelligence: Modeling the Decision Making of Human Operators in Highly Dangerous Situations

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

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

Abstract: A prescriptive simulation model of a process operator's decision making assisted with an artificial intelligence (AI) algorithm in a technical system control loop is proposed. Situations fraught with a catastrophic threat that may cause unacceptable damage were analyzed. The operators' decision making was interpreted in terms of a subjectively admissible probability of disaster and subjectively necessary reliability of its assessment, which reflect the individual psychological aspect of operator's trust in AI. Four extreme decision-making strategies corresponding to different ratios between the above variables were distinguished. 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 (its dynamics and the hazard onset's speed) 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; equipment predictive analytics

MSC: 91E45

check for updates

Citation: Venger, A.L.; Dozortsev, V.M. Trust in Artificial Intelligence: Modeling the Decision Making of Human Operators in Highly Dangerous Situations. *Mathematics* **2023**, *11*, 4956. https://doi.org/10.3390/math11244956

Academic Editors: Xiang Li, Natalia Bakhtadze, Igor Yadykin, Andrei Torgashov and Nikolay Korgin

Received: 23 September 2023
Revised: 2 December 2023
Accepted: 12 December 2023
Published: 14 December 2023

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 (HO), 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, explainability and ease-of-use of algorithms, responsibility for and benefits of the use of AI, etc.), an 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.

According to a review [1], the influence of personal characteristics on trust in AI depends on the significance of the decisions made with its assistance. It is also shown that, despite a large number of studies of the influence of an individual operator's personal characteristics on trust in AI, a comprehensive picture of this influence is currently missing.

The above problem was 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 in situations threatening unacceptable damage.

The structure of the work is as follows. Section 2 provides a brief overview of the problem of trust in technology in the context of the ever-increasing intellectualization of control systems for complex technical and technological objects. Section 3 discusses the requirements for human–machine interaction introduced by the advent of AI and known approaches to their implementation. The general task of preparing a HO for effective interaction with AI, including computer-based operator training, is formulated. In Section 4, a mathematical model of a human–machine system with an AI algorithm in the control loop is proposed, taking into account both the characteristics of the technical system, which are fundamental from the standpoint of trust in AI, and the strategies for a HO's decision making depending on their individual psychological characteristics (subjective anxiety). Section 5 presents the results of a large-scale simulation experiment with the proposed model, which allows for calculating the effectiveness of operator strategies depending on the characteristics of the controlled object and AI algorithms' parameters. Section 6 discusses the results of the experiment and its values for the development of a sound methodology for computer-based operator training, which allows us to form and optimize the skill of controlling a technical system in the presence of AI.

2. The Problem of Human Trust in Complex Machinery

The trust of a HO 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 Luddites were destroying power looms that were making British weavers' jobs redundant [2].

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 [3].

With advances in technology, the problem of trust has become increasingly multidimensional and philosophers, literary figures, sociologists, culturologists, psychologists, and science fiction writers are becoming involved in it. The adoption of computer-integrated manufacturing affects a new important category of users—industrial system operators.

The problem of trust in technology is a traditional topic for psychological research [4]. 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 HO profile: experience, professional competence, self-esteem, and other personal traits [5].

Nowadays, HOs must increasingly embrace recommender systems and intelligent decision support that have become indispensable for the 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 HO.

3. Specifics of Trust In/Distrust of AI Systems

The fundamental problem of HO–AI interaction is that most AI algorithms are based on machine learning (ML), and it is difficult for a HO to understand the logic of the solutions they offer. At the same time, the most effective algorithms (for example, deep learning) are often the least transparent, and vice versa (for example, decision trees) [6]. In modern production, HO–AI cooperation is accompanied by uncertainty and risk, and the operator's distrust of AI is an attempt to reduce these factors [7]. A so-called explainable artificial intelligence (XAI) should increase the trust of the end user by giving them explanations contextual to the subject area, capabilities and the user's expectations [6]. The relation of AI's clarity and predictability with trust in it (especially in potentially dangerous situations) is emphasized by many researchers [8–13].

A pragmatic view of AI in industrial automation suggests that the user attributes a human-like reasoning process to an AI [13]; therefore, it seems logical to apply theory of mind (ToM) approaches to the HO–AI interaction, making the mental state of interacting agents transparent, which is necessary for cognition and interaction [9]. An AI should explain its decision considering a human's understanding and intentions, as well as a person's understanding of AI [10].

With an abundance of works on the topic, specific models of trust in AI are rarely discussed [13], and the traditional design of HO–AI interaction is based on the external behavior of a HO, whose internal mental states are ignored, which potentially limits the effectiveness of the created systems [14].

Another developed area of research on user–AI interaction is the technology acceptance model (TAM). Originally proposed in the 1990s, it operates with two basic factors: perceived usability and perceived ease of use [15]. In the subsequent version (unified theory of acceptance and use of technology—UTAUT) [16], the factors of social influence (adjusting user's behavior to meet the demands of a social environment) and facilitating conditions (user believing in support of existing organizational and technical infrastructure) were added [17]. Recently, there have been publications on AI acceptance (AI-UTAUT); see review [18].

AI-UTAUT operates with some additional factors specific to AI and, at least indirectly, related to the user's trust in it—transparency (the AI decisions' clarity and understanding), explainability, anthropomorphism (attribution of human traits and intentions to AI) and value alignment (aligning AI's values and ethics with human values and social norms). It is emphasized that the acceptance of AI also depends on the specific AI algorithm that is being used [18].

With the obvious rationality of the AI-UTAUT approach, it should be noted that its applications to the industrial AI domain are still unknown, as are attempts to consider the user's individual psychological characteristics in the HO–AI interaction model. It is also wrong to neglect the accompanying, but no less important factors that aggravate the problem of trust in the AI.

Motivational trap. If correct AI advice is rejected or the wrong advice is accepted, the HO operator risks revealing a skills gap. If correct advice is accepted, then the AI will prove to be at least faster, or even “smarter” than the HO. Only rejecting the wrong advice would benefit a HO. Given an 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 eternal to a degree. An obvious tool to improve staff willingness is the training and re-profiling of users and professionals, among other things, with reliance on a technological facility with 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 the case of situations that were not covered during the algorithm’s learning stage;
- Staff fears of losing jobs are heightened at every critical point in the development of automation: during the 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 [19]. Now, automation claims not only fast routine tasks that are beyond human reach but also optimization, planning and predictive analytics, i.e., it increasingly encroaches on the “sancta sanctorum”, which is widely viewed as accessible to natural human intelligence only.

The control skill of a technical system containing an AI adviser can be attributed to the procedural skills of operators. The formatting, enhancing and improving of such skills are implemented by operator computer-based training with high-fidelity modeling of a technical system [20]. Similar examples of the inclusion of advanced automation systems in computer simulators are already available [21], and interest in them will increase as more and more intelligent control tools are transferred to the operator’s area of responsibility. However, there is no comprehensive methodology for training operators to interact with AI, so the general problem of training operational personnel to work with AI can be considered as the implementation of the following multi-step plan:

1. Build a model of the operator’s decision making (strategies for accepting/rejecting AI advice), considering, among other things, the individual psychological characteristics of HO;
2. Verify (in a simulation experiment) the proposed model’s rationality from the point of view of the significantly greater effectiveness of individual strategies in comparison with others in different operating conditions (including technological objects and the AI algorithm);
3. Study real subjects’ strategies in an engineering psychological experiment for their compliance with the proposed model and adaptation to changing operating conditions of the technical system;
4. Develop and verify the HO’s computer training methodology in the presence of AI in the control loop;
5. Develop and implement AI adaptation tools to the characteristics of the trained personnel (preferred strategies, the level of necessary information support, etc.);
6. Introduce the simulator at a large scale into the operator training process and analysis of its results.

This work is devoted to solving the first two subtasks, which are mathematical in nature and are a prerequisite for general problem solving.

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

The “Big Five” personality factors [22] that may influence trust in AI recommendations include extraversion, agreeableness, conscientiousness, neuroticism and openness to experience [1,23]. Thus, a positive influence on trust in AI has been shown for such traits as agreeableness [24,25], openness [26,27] and extraversion [28,29]. On the contrary, some studies demonstrated that neuroticism and anxiety have a negative impact on trust in AI [24,30–32]. We focused on the latter factor by the reason of strong influence of anxiety on high-risk decision making. Empirical studies have found that an increase in the level of anxiety corresponds with a tendency to avoid risk [33–36]. This corresponds to the approach to the problem of trust in the mentioned studies [24,30–32], where trust appears as a risky choice and distrust as a safer one. However, in the situation we are considering, when making a decision is impossible without interaction with AI, the picture changes. In this case, trusting the AI's warning of a threat is safer than ignoring such a warning.

We already proposed a generalized description of decision making in situations characterized by the risk of catastrophic consequences [37]. 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" [38] (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 determines decision making in high-risk situations to a very large extent [1,33–35] determining the tendency to choose less risky actions. According to our hypothesis, anxiety 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 themselves. "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 sequential decision theory [39], 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 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 modeled:

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

$$X_0 = 1/2\Delta,$$

where

Δ is the value, which if exceeded, is treated as a "catastrophe" resulting in unacceptable damage;
z_i~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, the AI also "knows" the predicted values of X'_{i+1} and X'_{i+2}; if $X'_{i+1} > \Delta$ or $X'_{i+2} > \Delta$, the 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}, that 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.

III. The model assumes that the operator is guided by both AI signals and their own assessment of the state of the process. They, however, are 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:

$$Y^{btm}{}_i = X_i + k \times (w_i - 2.5);$$

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

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].

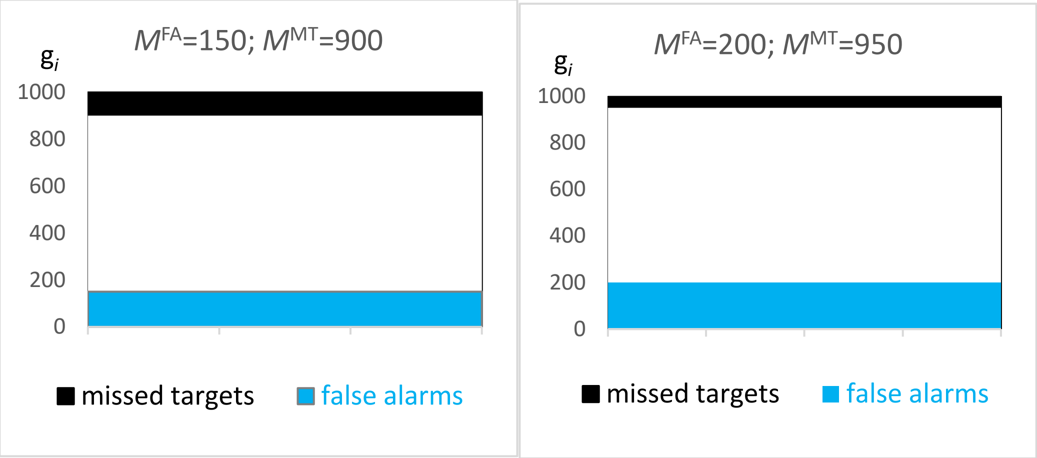

Figure 1. Zones of false alarms and missed targets at different values of M^{FA} and M^{MT}.

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 $0 < \delta^{frw} \leq \delta^{stp}$, which characterize the individual strategy (the higher the level of "apprehension", the lower δ^{stp} is; the higher the level of "doubt", the lower δ^{frw} is):

- If $Y_i \leq \delta^{frw}$, then a D^{frw} decision is made, i.e., the [0, δ^{frw}] segment is the *risk acceptance zone*;
- If $\delta^{frw} < Y_i \leq \delta^{stp}$, then a D^{tst} decision is made, i.e., the (δ^{frw}, δ^{stp}] interval is the *zone of uncertainty* (in case of $\delta^{stp} = \delta^{frw}$ it is absent);
- If $Y_i > \delta^{stp}$, 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., $X'_{i+1} > \Delta$ or $X'_{i+2} > \Delta$, 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: $X_{i+1} = 0$; $X_{i+2} = ½\Delta$.

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

Two "exploratory steps" in a row are impossible.

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.

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

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

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

where

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

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

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

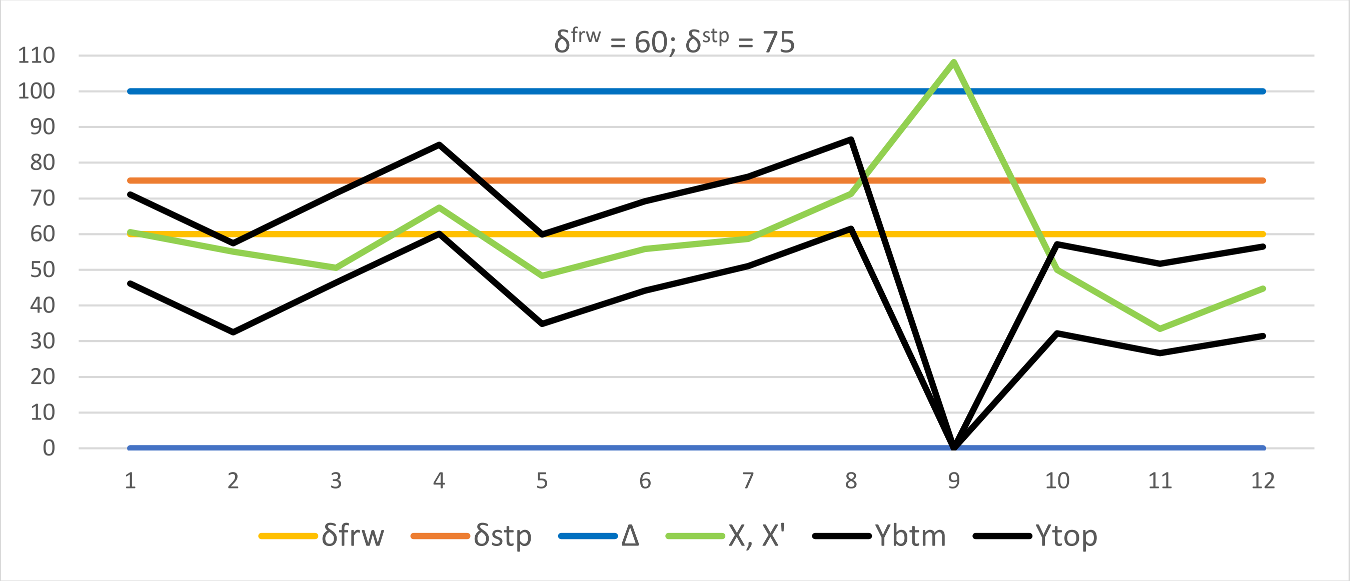

Figure 2. *X*-axis—step number, *Y*-axis—a subjective assessment of the state of the X_i. Alarm signal at step #7; $Y^{\text{btm}}{}_7 = 51.1$; $Y^{\text{top}}{}_7 = 76.1$; $Y_7 = ½ (Y^{\text{btm}}{}_7 + Y^{\text{top}}{}_7) = 63.6$; $\delta^{\text{frw}} < Y_7 < \delta^{\text{stp}}$ => D^{tst} decision; $Y^{\text{btm}}{}_8 = 61.5$; $Y^{\text{top}}{}_8 = 86.5$; $Y_8 = ½ (Y^{\text{btm}}{}_8 + Y^{\text{top}}{}_8) = 74.0$; $Y_8 > \delta^{\text{frw}}$ => D^{stp} final decision.

With full confidence that the 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 the AI ($h = 0$), the 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 an 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 them to make bolder decisions.

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):

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

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.

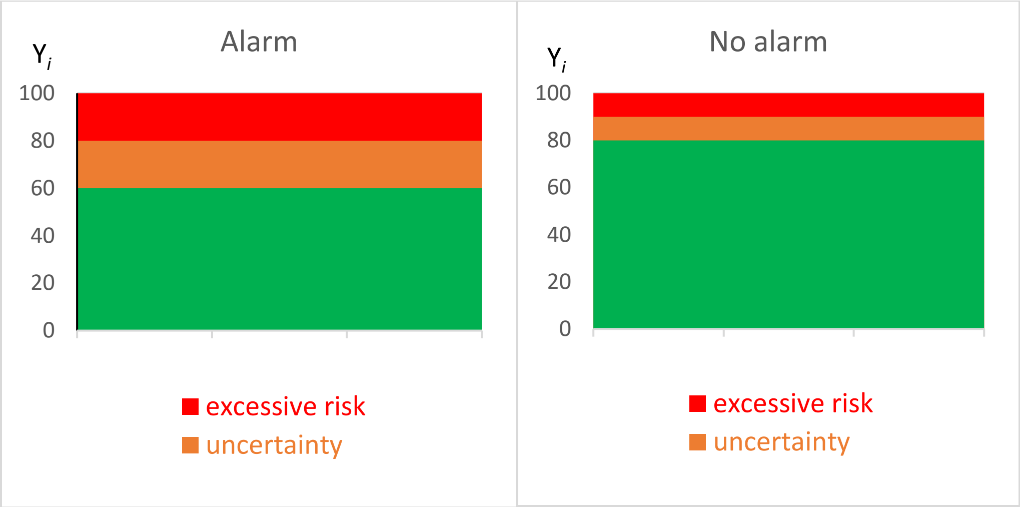

Figure 3. Zones of risk acceptance (green), uncertainty and excessive risk depending on the presence or absence of an alarm signal (δ^{frw} = 60, δ^{stp} = 80; h = 0.5; $\delta_h{}^{frw}$ = 80, $\delta_h{}^{stp}$ = 90). *Y*-axis—Y_i (a subjective assessment of the state of the X_i).

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

5. Outcomes

The model was tested to compare the predicted effectiveness of different operator strategies. Therefore, parameters δ^{frw} and δ^{stp} characterizing such strategies were varied.

In accordance with the research objective, when implementing the model, process options with varying degrees of risk were tested, i.e., frequency of emergency situations (parameter *b*). The presence/absence of AI "missing the target" (MMT), the accuracy of the subjective assessment of the process state (*k*) and the cost of verification (u^{tst}) also varied. The remaining parameters were represented by two contrasting values of *b* (with a low and high objective probability of process emergency states), two different values of M_{MT} (presence/absence of AI "missed targets") and two different values of *k* (the accuracy of the subjective assessment of the process state). 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. Δ = 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. (A high frequency of "false alarms" is important when training operators to recognize them and is, therefore, typical in process simulations.)
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 ($h = 1$), we predicted a low cost of verification ($u^{tst} = 1$).

2. Requiring much more caution were 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 ($M^{\mathrm{MT}} = 800$). Cases with a low verification cost ($\mathrm{u}^{\mathrm{tst}} = 1$; 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 1000 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 $\delta^{\mathrm{frw}} = 55$ since the initial state of the process was $\mathrm{X}_0 = 50$. The values of $\delta^{\mathrm{frw}} = 55$ and $\delta^{\mathrm{frw}} = 60$ were considered low (*risk avoidance* strategies), $\delta^{\mathrm{frw}} = 65$ and $\delta^{\mathrm{frw}} = 70$ were considered medium, and $\delta^{\mathrm{frw}} = 75$ 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 a 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} [40]. 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 was 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 $\delta^{\mathrm{frw}} = 75$. At higher values of δ^{frw}, the average frequency of "accidents" exceeded the maximum permissible level. So, already at $\delta^{\mathrm{frw}} = \delta^{\mathrm{stp}} = 80$, it was 0.080 per 1000 steps. At $\delta^{\mathrm{frw}} < 75$, the integral indicator U significantly decreased. Thus, at $\delta^{\mathrm{frw}} = 70$, the maximum result reaches 19,759 (at $\delta^{\mathrm{stp}} = 85$), which is significantly lower than the result in the pair ($\delta^{\mathrm{frw}} = 75$, $\delta^{\mathrm{stp}} = 85$), which is 20,242 ($p < 0.001$; hereinafter, the significance of the differences was checked using 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 $\delta^{\mathrm{frw}} = 75$ and variable values of δ^{stp} are set out in Table 1.

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

		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	1035	1027	1002	975	1037

As it appears from the Table, the optimal combination was ($\delta^{\mathrm{frw}} = 75$, $\delta^{\mathrm{stp}} = 85$). The difference between the integral indicator U in pairs ($\delta^{\mathrm{frw}} = 75$, $\delta^{\mathrm{stp}} = 85$) and ($\delta^{\mathrm{frw}} = 75$, $\delta^{\mathrm{stp}} = 75$) 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).

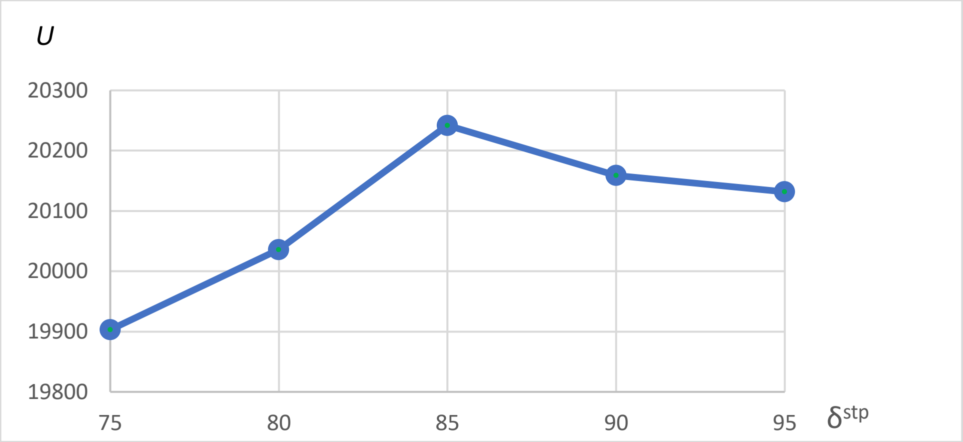

Figure 4. Integral indicator U at $\delta^{frw} = 75$ depending on δ^{stp}.

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.

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	2436	2082	2202	1893	2182	1693	1648	1880
U_2	M	11,293	12,627	13,152	13,151	13,914	14,867	15,017	15,678
	S	2436	2186	2452	2169	2182	1781	1832	1880

As 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).

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").

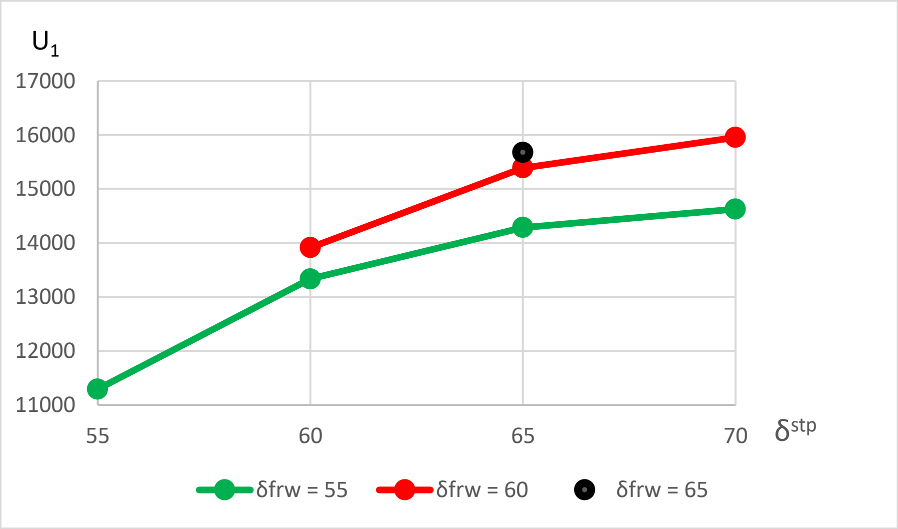

Figure 5. Integral indicator U_1 at different values of δ^{frw} and δ^{stp}.

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 ($\delta^{\mathrm{frw}} = \delta^{\mathrm{stp}} = 65$), which provides a significantly higher U_2 than each of the other reviewed scenarios ($p < 0.001$) (Figure 6).

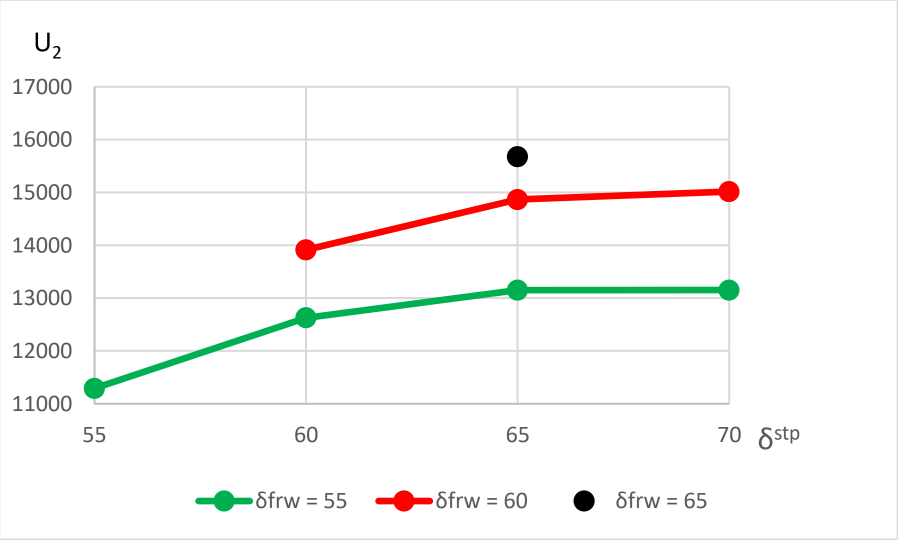

Figure 6. Integral indicator U_2 at different values of δ^{frw} and δ^{stp}.

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 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 [41]. 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 the AI recommendations.

The described model represents a simple random dynamic process at risk of a critical regime violation. It is assumed that the AI sometimes signals to a human operator that such a deviation is approaching and recommends stopping the process. According to the model, the operator accepts/checks/rejects AI advice in accordance with the values of two key characteristics of anxiety. It is shown that for different process parameters and AI algorithms, the effectiveness of decision-making strategies varies significantly. This result is a necessary, but insufficient premise for transferring such an approach to an operator's training to effectively manage the industrial process with AI assistance.

The next stage of the experiment will involve human subjects to decide how to use AI advice in dangerous situations based on their individual decision-making strategies. The anxiety parameters and, consequently, the subjects' strategies, will be evaluated according to whether real decision-making points belong to "risk acceptance", "uncertainty" or "excessive risk" zones. The limitations of this approach relate to the stability of strategies and their adaptability through special training when conditions change. This training could be implemented using computer-based simulators, based on high-fidelity process modeling, and providing a high-level psychological similarity of the operator's activity in the training and at the workplace [20]. Such simulators already include intelligent process control systems, for example, based on predictive models [21]. The value of the proposed model also lies in the fact that it can be used to adapt AI algorithms to the individual characteristics of operators' anxiety (choosing the necessary level of transparency and explainability of AI advice), which is also vitally important for increasing operator's trust in artificial intelligence.

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. Operator reluctance to use AI is largely driven by an 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 AI system skills through training and re-training on the basis of proven computer simulators, including a high-fidelity model of a 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.

Author Contributions: Conceptualization, A.L.V. and V.M.D.; Methodology, A.L.V. and V.M.D.; Formal analysis, A.L.V.; Investigation, V.M.D.; Writing—original draft, A.L.V. and V.M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Riedl, R. Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electron. Mark.* **2022**, *32*, 2021–2051. [CrossRef]
2. Jones, S.E. *Against Technology: From the Luddites to Neo-Luddism*; Taylor & Francis: Oxford, UK, 2006.
3. Hart, G.; Goldwater, B. *Recent False Alerts from the Nation's Missile Attack Warning System*; U.S. Government Printing Office: Washington, DC, USA, 1980.
4. Lee, J.; See, K. Trust in technology: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. [CrossRef]
5. Akimova, A.; Oboznov, A. The factors of increase in trust and decrease in distrust of human to technique. *Psychol. Stud.* **2017**, *10*, 8. [CrossRef]
6. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI-Explainable artificial intelligence. *Sci Robot.* **2019**, *4*, eaay7120. [CrossRef]
7. Hoffman, R.R. A taxonomy of emergent trusting in the human–machine relationship. In *Cognitive Systems Engineering: The Future for a Changing World*; CRC Press: Boca Raton, FL, USA, 2017; pp. 137–163.
8. Alonso, V.; De La Puente, P. System Transparency in Shared Autonomy: A Mini Review. *Front. Neurorobot.* **2018**, *12*, 83. [CrossRef]
9. Williams, J.; Fiore, S.M.; Jentsch, F. Supporting Artificial Social Intelligence with Theory of Mind. *Front. Artif. Intell.* **2022**, *5*, 750763. [CrossRef]
10. Akula, A.R.; Liu Sari Ch Saba-Sadiya, S.; Lu, H.; Todorovic, S.; Chai, J.Y.; Zhu, S.C. X-tom: Explaining with theory-of-mind for gaining justified human trust. *arXiv* **2019**, arXiv:1909.06907.
11. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *46*, 1–38. [CrossRef]
12. Papagni, G.; Koeszegi, S. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn J. Behav. Robot.* **2020**, *12*, 13–30. [CrossRef]
13. Jacovi, A.; Marasović, A.; Miller, T.; Goldberg, Y. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual, 3–10 March 2021; pp. 624–635.
14. Oguntola, I.; Hughes, D.; Sycara, K. Deep interpretable models of theory of mind. In Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), Vancouver, BC, Canada, 8–12 August 2021; pp. 657–664.
15. Adams, D.A.; Nelson, R.R.; Todd, P.A. Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS Q.* **1992**, *16*, 227–247. [CrossRef]
16. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User acceptance of information technology: Toward a unified view. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]
17. Li, Y.; Zhao, M. A Study on the Influencing Factors of Continued Intention to Use MOOCs: UTAUT Model and CCC Moderating Effect. *Front. Psychol. Sec. Educ. Psychol.* **2021**, *12*, 528259. [CrossRef]
18. Fetaji, M. Devising a Model AI-UTAUT by Combining Artificial Intelligence (AI) with Unified Theory of Acceptance and Use of Technology (UTAUT). *SAR J.* **2023**, *6*, 182–187. [CrossRef]

19. Alekseev, A.; Garbuk, S. How can you trust Artificial Intelligence Systems? Objective, Subjective and Intersubjective parameters of Trust. *Artif. Soc.* **2022**, *17*, 2. [CrossRef]
20. Dozortsev, V.M.; Agafonov, D.V.; Nazin, V.A.; Novichkov, A.Y.; Frolov, A.I. Computerized operator training: Continued importance, new opportunities, and the human factor. *Autom. Remote Control* **2020**, *81*, 935–954. [CrossRef]
21. Toro, R.; Ortiz, J.M.; Yutronic, I. An Operator Training Simulator System for MMM Comminution and Classification Circuits. In Proceedings of the IFAC Workshop on Automation in the Mining, Mineral and Metal Industries, Gifu, Japan, 10–12 September 2012.
22. John, O.P.; Naumann, L.P.; Soto, C.J. Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In *Handbook of Personality: Theory and Research*, 3rd ed.; John, O.P., Robins, R.W., Pervin, L.A., Eds.; The Guilford Press: New York, NY, USA, 2008; pp. 114–158.
23. Matthews, G.; Hancock, P.A.; Lin, J.; Panganiban, A.R.; Reinerman-Jones, L.E.; Szalma, J.L.; Wohleber, R.W. Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personal. Individ. Differ.* **2021**, *169*, 109969. [CrossRef]
24. Kraus, J.; Scholz, D.; Baumann, M. What's driving me? Exploration and validation of a hierarchical personality model for trust in automated driving. *Hum. Factors* **2020**, *63*, 1076–1105. [CrossRef]
25. Rossi, S.; Conti, D.; Garramone, F.; Santangelo, G.; Staffa, M.; Varrasi, S.; Di Nuovo, A. The role of personality factors and empathy in the acceptance and performance of a social robot for psychometric evaluations. *Robotics* **2020**, *9*, 39. [CrossRef]
26. Antes, A.L.; Burrous, S.; Sisk, B.A.; Schuelke, M.J.; Keune, J.D.; DuBois, J.M. Exploring perceptions of healthcare technologies enabled by artificial intelligence: An online, scenario-based survey. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 221. [CrossRef] [PubMed]
27. Oksanen, A.; Savela, N.; Latikka, R.; Koivula, A. Trust toward robots and artificial intelligence: An experimental approach to human–technology interactions online. *Front. Psychol.* **2020**, *11*, 568256. [CrossRef] [PubMed]
28. Böckle, M.; Yeboah-Antwi, K.; Kouris, I. Can you trust the black box? The effect of personality traits on trust in AI-enabled user interfaces. In *Artificial Intelligence in HCI*; Degen, H., Ntoa, S., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2022; Volume 12797, pp. 3–20. [CrossRef]
29. Haring, K.S.; Matsumoto, Y.; Watanabe, K. How do people perceive and trust a lifelike robot. In Proceedings of the World Congress on Engineering and Computer Science, San Francisco, CA, USA, 23–25 October 2013; pp. 425–430.
30. Miller, L.; Kraus, J.; Babel, F.; Baumann, M. More than a feeling—Interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Front. Psychol.* **2021**, *12*, 592711. [CrossRef]
31. Dekkal, M.; Arcand, M.; Prom Tep, S.; Rajaobelina, L.; Ricard, L. Factors affecting user trust and intention in adopting chatbots: The moderating role of technology anxiety in insurtech. *J. Financ. Serv. Mark.* **2023**, 1–30. [CrossRef]
32. Zhang, T.; Tao, D.; Qu, X.; Zhang, X.; Zeng, J.; Zhu, H.; Zhu, H. Automated vehicle acceptance in China: Social influence and initial trust are key determinants. *Transp. Res. Part C Emerg. Technol.* **2020**, *112*, 220–233. [CrossRef]
33. Maner, J.; Schmidt, N. The role of risk avoidance in anxiety. *Behav. Ther.* **2006**, *37*, 181–189. [CrossRef]
34. Maner, J.K.; Richey, J.A.; Cromer, K.; Mallott, M.; Lejuez, C.W.; Joiner, T.E.; Schmidt, N.B. Dispositional anxiety and risk-avoidant decision-making. *Personal. Individ. Differ.* **2007**, *42*, 665–675. [CrossRef]
35. Hengen, K.M.; Alpers, G.W. Stress Makes the Difference: Social Stress and Social Anxiety in Decision-Making Under Uncertainty. *Front. Psychol. Sec. Decis. Neurosci.* **2021**, *12*, 578293. [CrossRef]
36. Charpentier, C.J.; Aylward, J.; Roiser, J.P.; Robinson, O.J. Enhanced Risk Aversion, But Not Loss Aversion, in Unmedicated Pathological Anxiety. *Biol. Psychiatry* **2017**, *81*, 1014–1022. [CrossRef]
37. Venger, A.L. Mathematical model of decision making in extreme situations. *Autom. Ind.* **2018**, *6*, 32–36.
38. Lu, J.; Jain, L.C.; Zhang, G. Risk management in decision making. In *Handbook on Decision Making: V.2: Risk Management in Decision Making*; Springer: Berlin/Heidelberg, Germany, 2012.
39. LaValle, S.M. Sequential Decision Theory. In *Planning Algorithms*; Cambridge University Press: Cambridge, UK, 2006; Chapter 10; pp. 495–559.
40. Sim, K.B.; Lee, M.L.; Wong, S.Y. A review of landslide acceptable risk and tolerable risk. *Geoenviron. Disasters* **2022**, *9*, 3. [CrossRef]
41. 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. *Int. J. Sport Exerc. Psychol.* **2017**, *15*, 41–63. [CrossRef]

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.